ASSESSING SPECIES SPECIFIC DETERMINANTS OF DNA BINDING IN MCRB HOMOLOGS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Chris John Hosford

May 2019

ASSESSING SPECIES SPECIFIC DETERMINANTS OF DNA BINDING IN MCRB HOMOLOGS

Chris John Hosford, Ph. D.

Cornell University 2019

Antibiotic resistance poses a significant threat to human health. The lagging development of new antibiotics and the rapid exchange of resistance genes have created a need for alternative methods to combat emerging 'superbugs'. One promising strategy involves using lytic phages – bacterial viruses that lyse and kill their hosts, releasing their progeny. Modification-dependent restriction (MDR) systems are innate bacterial defense systems that recognize and degrade phage DNA, thus preventing mature phage production. MDR systems are highly conserved in antibiotic resistant bacteria and several phage-encoded inhibitors have been identified. Understanding the molecular mechanisms of these systems can therefore provide a means to increase the therapeutic potential of phages.

McrBC is a two-component MDR system that restricts DNA containing 4-, 5-, or 5-hydroxy-methyl cytosines. The first component, McrB, contains an N-terminal domain that recognizes and binds the modified site and a C-terminal AAA+ motor domain that hydrolyzes GTP and mediates nucleotide-dependent oligomerization. Bioinformatics coupled with a structure of the *Escherichia coli* McrB suggest that McrB homologs may target different nucleic acids using different molecular mechanisms. To assess the species-specific determinants of McrB DNA-binding, I have purified the

putative DNA binding domains of different McrB homologs and determined their atomic resolution structures by x-ray crystallography.

The structures of the *Thermococcus gammatolerans* (Tg) and *Staphylothermus marinus* McrB N-terminal domains were solved to 1.68 Å and 2.10 Å respectively and revealed structural homology to the PUA-like domains of RNA binding proteins. The structures identified a conserved aromatic cage required for RNA binding via base-flipping of a modified adenosine base. A structure of the TgMcrB in complex with DNA containing 5-methylcytosine confirms this base-flipping mechanism and provides a model for how these proteins bind modified nucleic acids. Furthermore, the structure of the *Helicobactor pylori* (Hp) LlaJI N-terminal domain was solved to 1.97 Å and revealed structural homology to the B3 family of site-specific DNA binding proteins. LlaJI is a homolog of McrB and functions as a restriction modification (R/M) system that targets DNA site specifically. Together these structures underscore the inherent structural plasticity of McrB DNA binding and provides insights into their molecular mechanisms of target specificity.

BIOGRAPHICAL SKETCH

My undergraduate work was completed in 2013 at the Georgia Institute of Technology prior to joining the department of Biochemistry, Molecular and Cell Biology at Cornell University. I joined the lab of Joshua Chappie in May of 2014 and have been a member in his lab until my commencement in May of 2019.

ACKNOWLEDGMENTS

This work could not have been accomplished without the guidance and support of many people who have helped me along the way:

To my advisor, Joshua Chappie, who has been incredibly supportive of all my endeavors, both in and outside of the laboratory.

To my beautiful wife, Jane, to always provide love, encouragement, and support when I needed it most.

To my parents, John and Hiroko Hosford, for all their support throughout the years. I would not be where I am without them.

To my sisters, Cynthia and Caitlin, for their endless confidence that I had what it took to make it in the end.

To the Chappie lab members, Carl Schiltz and Myfanwy Adams, and my committee members, Eric Alani and Chris Fromme, for their valuable insight, suggestions, and support.

To my dog, Luke, for his brilliant smile (and furry tail) and always giving me a reason to look forward to going home every day.

TABLE OF CONTENTS

Chapter 1. Introduction

In 1915, Frederick Twort unsuccessfully attempted to propagate vaccinia virus. Instead, he observed 'transparent' plaques on his plates, which upon closer examination, turned out to be zones of dead bacteria. Twort proposed three possible hypotheses to his observation: 1) it could be a novel manifestation of the bacterial life cycle, 2) an unknown protein or enzyme could be produced by the bacteria itself, or 3) it could be some sort of "ultra-microscopic virus" that infects bacteria (Keen et al., 2015). At the time, Twort's hypotheses could not be confirmed nor rejected. It wasn't until two years later that Felix d'Herelle published similar observations and attributed them to a new type of virus that infected bacteria which he later dubbed the bacteriophage (phage). Over the next century, phages would play a leading role in numerous biological advancements including discovery of DNA as the genetic material and how it is encoded as a triplet (codons).

By 1919, d'Herelle had speculated that phages were responsible for the recovery of a variety of illnesses that he proposed to employ laboratory-produced phage as both prophylactic and therapeutic agents against bacterial infection. The idea was to exploit the ability of phage to serve as 'bacterial-killers' and stave off infection. Due to its narrow specificity of cellular target hosts, phage therapy was initially only used to treat acute and chronic infections. Indeed, the only uses for phage therapy in the 1920 – 1930s was serum therapy against pathogens such as pneumococci and diphtheria (Wittebole et al., 2014). Unfortunately, skepticism and controversy surrounded phage therapy from the beginning due to early studies lacking appropriate controls and producing inconsistent results. Moreover, the emergence of penicillin as a 'broad spectrum' antibiotic in 1942 further dampened any interest in phage research and therapy

(Summers et al., 2012). However, over the last decade, the emergence of multi-drug resistant bacteria has led researchers to reconsider phage therapy as a viable alternative to antibiotics.

Phage virions are remarkably diverse and vary widely in size, shape, and complexity. Their genomes are even more diverse ranging in size from 3.4 kb to <500 kb and encode all the necessary components to successfully propagate the phage through its host. All phages can undergo the lytic cycle while only temperate phages exhibit bimodality and can form stable lysogens. The lytic state is a productive phase leading up to the synthesis of new phage particles. In the lytic cycle, the host cell machinery expresses phage genes, replicates the phage genome, and manufactures more phage particles to eventually rupture (lyse) the host, thereby killing the host and completing its reproductive cycle. The lysogenic state, however, is a dormant, or 'silent', phase where the viral genome is integrated within the host chromosome as a prophage. A complex regulatory network, best characterized in λ phage, regulates lytic gene inhibition and activation as a binary switch to convert it from the lysogenic state to the lytic state (Oppenheim et al., 2005).

The inherent nature of the lytic cycle leading to host cell death resulted in phages playing a significant role in driving bacterial evolution. Bacteria pose several layers of defense against phage infection, ranging from preventing phage adsorption to a wide arsenal of restriction systems poised to target and degrade the invading genetic material. The first layer of defense, preventing phage adsorption, is primarily extracellular and can be divided into three categories: blocking phage receptors, producing a complex extracellular matrix that can act as a physical barrier between phage and their receptors,

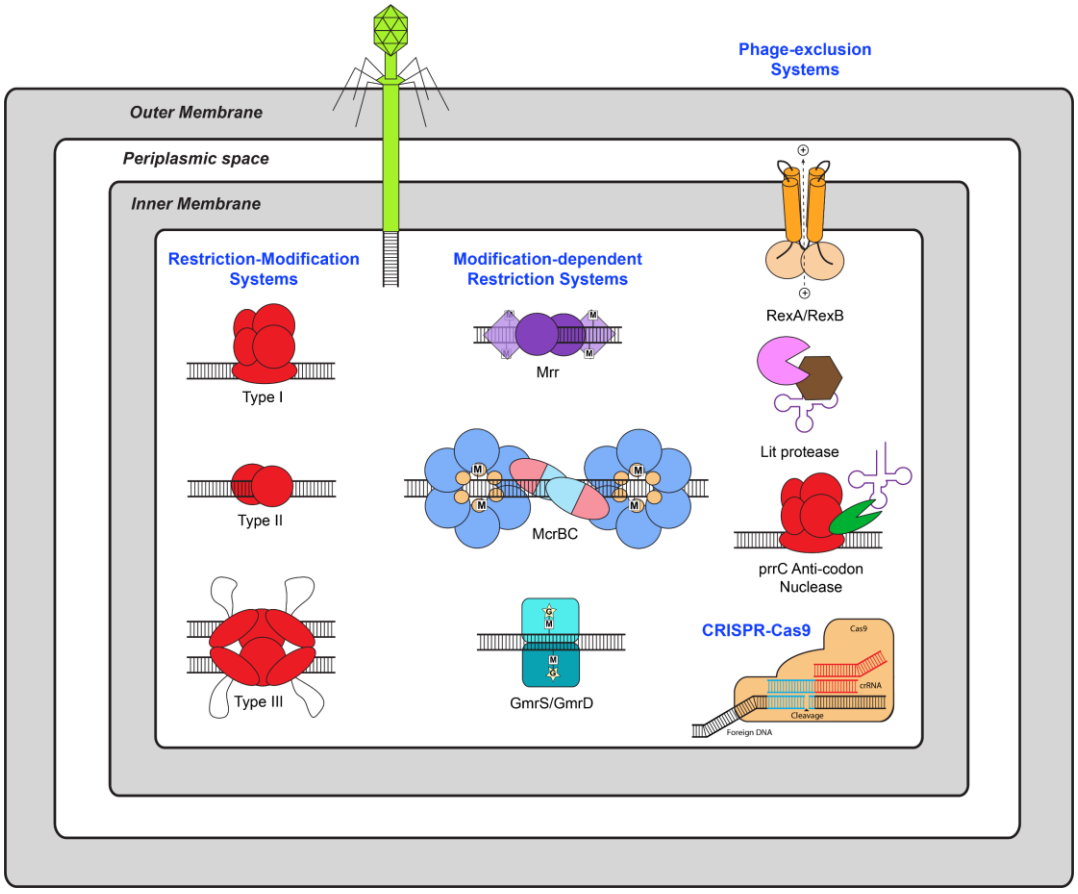

Figure 1. Overview of bacterial defense and phage exclusion systems. Restriction modification systems target DNA site specifically and include Types I, II, and III. Modification-dependent Restriction Systems target DNA modification specifically and include restriction enzymes like Mrr, McrBC, and GmrS/GmrD. Phage exclusion systems are bacterial suicide systems induced by phage encoded proteins and include examples like RexA, the Lit protease, and the prrC anticodon nuclease. CRISPR-Cas is a bacterial adaptive immune system that stores foreign DNA within the genome as a 'memory' to selectively target the same DNA in future infections.

or the production of competitive inhibitors against phage receptors. The second layer of defense resides intercellularly and involves the use of restriction enzymes (REases), or enzymes designed to target and cleave phage (and other foreign) nucleic acids (Labrie et al., 2010). Phage has in turn adapted chemical modifications to their genetic material to circumvent these systems. This development has resulted in an ongoing evolutionary arms race between bacteria and phage with bacteria evolving novel systems to target and degrade the adapted phage genome. Meanwhile, phage escape these systems by

incorporating different sequences, modifications, and in some cases, phage encoded inhibitors to their genome. Figure 1 briefly illustrates several mechanisms of bacterial defense systems including restriction modification (RM), modification dependent restriction (MDR), phage exclusion, and CRISPR-Cas systems. Each of these will be discussed in detail below.

RESTRICTION MODIFICATION SYSTEMS

Classical RM systems are site specific restriction systems that target and cleave (phage) DNA sequence specifically. The operons containing RM systems ubiquitously encode for an associated methyltransferase (MTase) designed to methylate the target site within the host genome to protect it from self-cleavage. Three classes of RM systems have been identified, Type I, II, and III, and vary in their structural composition, target recognition, and mechanism of restriction. Type I REases are large, heteropentameric proteins with separate restriction (R), methylation (M), and sequence recognition (S) subunits encoded by the host specificity determinant (*hsd)* genes, *hsdR*, *hsdM*, and *hsdS* respectively. The cleavage competent type I system is composed of R_2M_2S subunits and requires ATP, Mg^{2+}, and S-adenosylmethionine (SAM) for REase and MTase activity (Loenen et al., 2014). These systems exhibit bipartite DNA recognition through the S subunit with each assembly spaced as far as 2 kilobases (kb) apart. The R subunit is essential for REase activity, consisting of an N-terminal endonuclease fused to a C-terminal, RecA-like ATP-dependent motor domain (Iyer et al., 2004). ATP-dependent DNA translocation of type I REases has been observed to generate DNA loops visible by electron microscopy (EM) (Rosamond et al., 1979; Yuan et al., 1980) and atomic

force microscopy (AFM) (van Noort et al., 2004; Neaves et al., 2009). DNA is cleaved through the R subunits once translocation is blocked, either by collision with another molecule or by the presence of supercoiled DNA (Loenen et al., 2014). Examples of Type I RM systems include the *Escherichia coli* (*E. coli*) *Eco*KI and *Eco*BI.

Unlike type I RM systems, type II are enormously useful REases in the enzymatic toolboxes of molecular biologists. Type II REases are smaller, homodimeric proteins consisting of R_2 subunits capable of target recognition and cleavage accompanied by a separate M subunit for methylase activity. Alternative fused forms of R and M (R~M) have been observed and can form homodimeric $(R{\sim}M)_2$ complexes. These systems target short, symmetric (palindromic) sequences site-specifically and can readily cleave DNA in the absence of ATP. The first type II REase discovered was *Haemophilus influenza,* serotype d (*Hind*III, AAGCTT) by Hamilton Smith and Daniel Nathans which was awarded the Nobel Prize in Physiology or Medicine in 1978. This discovery spurred a growing interest in restriction enzymes eventually resulting in the discovery of two of the best characterized type II REases, *Eco*RI (GAATTC) and *Eco*RV (GATATC). Crystal structures of *Eco*RI and *Eco*RV in complex with their cognate DNA sequence yielded significant insight into the biomolecular mechanism of DNA recognition and cleavage. The first of these, *Eco*RI, was reported in 1986 and was crystallized with self-complementary 12- and 13-mer oligos in the absence of Mg^{2+} to avoid DNA cleavage (McClarin et al., 1986). The second, *Eco*RV, was reported seven years later in both a DNA-bound and unbound state, also in the absence of Mg^{2+} (Winkler et al., 1993). These two structures allowed for several key generalizations to be made about type II REases:

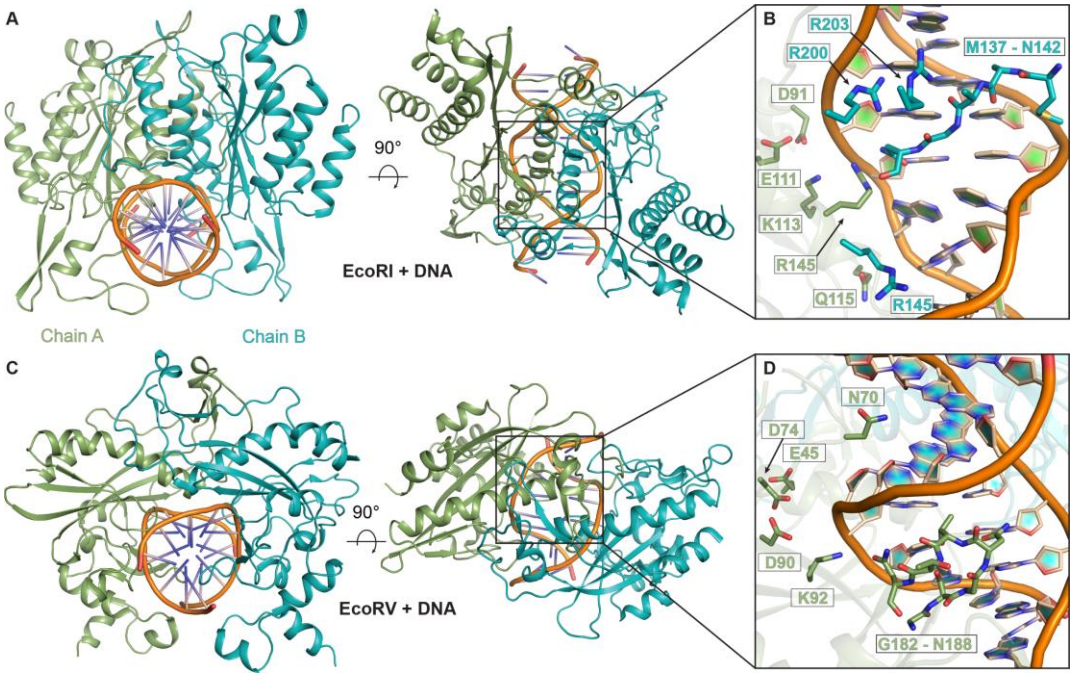

Figure 2. Structural comparison of *Eco*RI and *Eco*RV bound to their cognate DNA sequences. *A*. Cartoon representation of *Eco*RI in complex with DNA containing its cognate GAATTC sequence showin in 2 orientations. *B*. Zoomed in view of the *Eco*RI DNA binding and cleavage sites. D91, E111, and K113 form the conserved PD-(DE)XK motif. The remaining residues are involved in site-specific recognition. *C*. Cartoon representation of *Eco*RV in complex with DNA containing its cognate GATATC sequence showin in 2 orientations. *D*. Zoomed in view of the *Eco*RV DNA binding and cleavage sites. D74, D90, and K92 form the conserved PD-(DE)XK motif. The remaining residues are involved in site-specific recognition.

1. They form symmetric homodimers (2R) that are coincident with their palindromic recognition sequence (Figures 2A and 2C). The recognition and cleavage machinery from both subunits are oriented equally around each half-site in the palindromic sequence.
2. The substrate DNA is bound in a high energy conformation (underwound in *Eco*RI and kinked in *Eco*RV). This distortion is part of the recognition process and is coupled to conformational changes in the protein.
3. They both contain a central, four-stranded β-sheet flanked by two α-helices on both sides (forming an αββββαβ topology). This core fold was subsequently found (with variations) in almost all Type II REases whose structures have been determined. It

is characterized in the SCOP (**s**tructural **c**haracterization **o**f **p**roteins) database as the 'REase-like fold'.

4. Site-specific DNA recognition is achieved by a redundant network of extensive contacts in the major groove (Figures 2B and 2D). This ensures that contacts to the base pairs are over-determined, thereby increasing reliability.
5. The catalytic site consists of two acidic residues and one basic residue, falling under the canonical PD-(D/E)XK family of endonucleases (Figures 2B and 2D). Previous bioinformatic analysis indicated that among 289 Type II REases, 69% belonged to the PD-(D/E)XK phosphodiesterase superfamily (Orlowski et al., 2008).

Most of these residues in *Eco*RI and *Eco*RV were probed by site-directed mutagenesis experiments which have confirmed their importance in DNA binding or cleavage. Since then, over 30 additional Type II REase – DNA complexes have been solved at increasingly higher resolution and provide only a snapshot of the dynamic process of DNA cleavage by these enzymes.

Although the generalizations to Type II REases by structural data like *Eco*RI and *Eco*RV are largely consistent across the family, as more enzymes are discovered, more marked differences among them became apparent. Even amongst enzymes with comparable activities, little similarity was found at the amino acid sequence level. This diversity came as a surprise to many investigators in the field, and to date, there is still no consensus on what it means evolutionarily. Thus, characterizing Type II variants by the conventional genotype grouping was impractical. Instead, Type II REases were grouped by phenotype based on their behavior and cleavage properties, a scheme proposed by Roberts over a decade ago (Roberts et al., 2003). *Eco*RI, *Eco*RV, and most

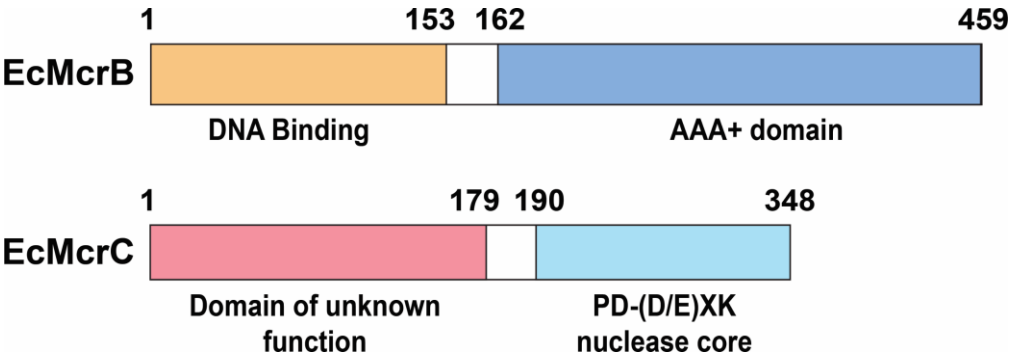

Figure 3. Domain architecture of McrBC. *Top*. Domain architecture of EcMcrB illustrating the N-terminal putative DNA binding domain and C-terminal AAA+ domain. *Bottom*. Domain architecture of EcMcrC illustrating the N-terminal DUF and C-terminal canonical PD-(D/E)XK endonuclease.

of the familiar laboratory cloning enzymes belong to the Type IIP subtype because they recognize **p**alindromic DNA sequences. Additional subtypes include Type IIL, a unique R~M REase that cleaves and methylates in one bi-functional active site (e.g. *Mme*I) (Callahan et al., 2016) and Type IIM, which includes REases capable of cleaving modified DNA (e.g. *Dpn*I and MspJI) (Siwek et al., 2012; Horton et al., 2014). Although many other Type II variants exist, there are far too many to describe each in detail.

Type III RM systems are perhaps the least well-characterized with only ~140 confirmed and putative types discovered. These systems also comprise of separate R and M subunits forming homodimeric M_2 and heterotetrameric R_2M_2 complexes. The M subunit in M_2 or R_2M_2 is responsible for DNA methylation while the R subunit is responsible for ATP hydrolysis, DNA translocation, and cleavage (Raghavendra et al., 2012). They are like Type I systems as they require ATP-dependent translocation for long-distance DNA cleavage, however, differ in that they require two inversely oriented recognition sites that can vary in their spatial orientation. (Meisel et al., 1992). Examples of Type III RM systems include the *E. coli Eco*P1I and *Eco*P15I.

MODIFICATION DEPENDENT RESTRICTION SYSTEMS

Regardless of their structural and biochemical differences, all Type I – III RM (excluding Type IIM) systems share at least one key feature: they target DNA site-specifically and are protected against by site-specific modification. Biological DNA modifications have been known for many years to play critical roles in all domains of life, including eukaryotes, prokaryotes, and bacteriophages. In eukaryotes, the modification N6-methyladenine (m^6A) has been shown as a major cancer cell marker and its abundance is directly linked to tumorigenesis (Xiao et al., 2018) while 5-methylcytosine (m^5C) modifications in CpG islands plays a major regulatory role in transcriptional gene activation and silencing (Handy et al., 2011). Like eukaryotes, bacteria also make widespread use of post-replicative DNA methylation for epigenetic regulation. However, instead of cytosine methylation, bacteria primarily make use of adenine methylation. Furthermore, bacteria and their predatory bacteriophage also have the added benefit of DNA modifications to protect their genetic material from RM systems. These modifications are the quintessential driver in the evolutionary arms race between bacteria and phage and led to the development of MDR systems to protect the host against the modified phage genomes.

Modification-dependent restriction was first observed in 1952 with T4 phage that contained hydroxymethylcytosine (hm^5C)-substituted DNA and marks the discovery of the first restriction system by Luria and Human (Luria and Human, 1952). The original observation led to the discovery of modifying enzymes that glucosylate hm^5C in T-even phages and of genes encoding enzymes that restrict non-glucosylated phage DNA: *rglA and rglB* (restricts glucoseless phage). These genes were later

renamed to *mcrA and mcrBC* (**m**odified **c**ytosine **r**estriction) and formed the new class of MDR systems, or Type IV systems, that recognize and cleave modified DNA. Unlike classical RM systems that contain an associated MTase to block cleavage, operons encoding for MDR systems only include genes for site recognition and restriction. McrA and McrBC are prototypical members and target DNA containing 4-methylcytosine (m^4C), m^5C, and hm^5C. Together, McrA and McrBC form the *E. coli* immigration control system for modified DNA. A third *E. coli* enzyme, *mrr* (**m**odified DNA **r**ejection and **r**estriction), targets DNA containing either methylcytosine or methyladenine. Since their discovery, additional key MDR systems have been identified including PvuRts1I, MspJI (from Type IIM), and GmrSD.

McrA was originally identified in *E. coli* K-12 (*Eco*K12) as a REase that restricted non-glucosylated, hm^5C-containing T-even phages, a group of dsDNA bacteriophages from *Myoviridae* that infect *E. coli*. It was subsequently demonstrated to restrict m^5C modified DNA in a Mn^{2+} and sequence dependent manner. For instance, McrA has been shown to be very effective in restricting DNA that has been methylated by M.*Hpa*II (Cm^5CGG) but not by other methyltransferases (Raleigh et al., 1986). Electrophoretic mobility shift assays (EMSAs) suggest that the true specificity is Ym^5CGR where Y is any pyrmidine (T or C) and R is any purine (A or G) (Mulligan et al., 2010). A recent crystal structure in the absence of DNA revealed that *Eco*K12 McrA dimerizes through its C-terminal HNH endonuclease domain whereas its N-terminal domain contains a putative modification-dependent binding domain (Czapinska et al., 2018). The overall domain organization is reminiscent of the two-domain organization of the SRA-HNH endonuclease previously characterized where DNA binding is

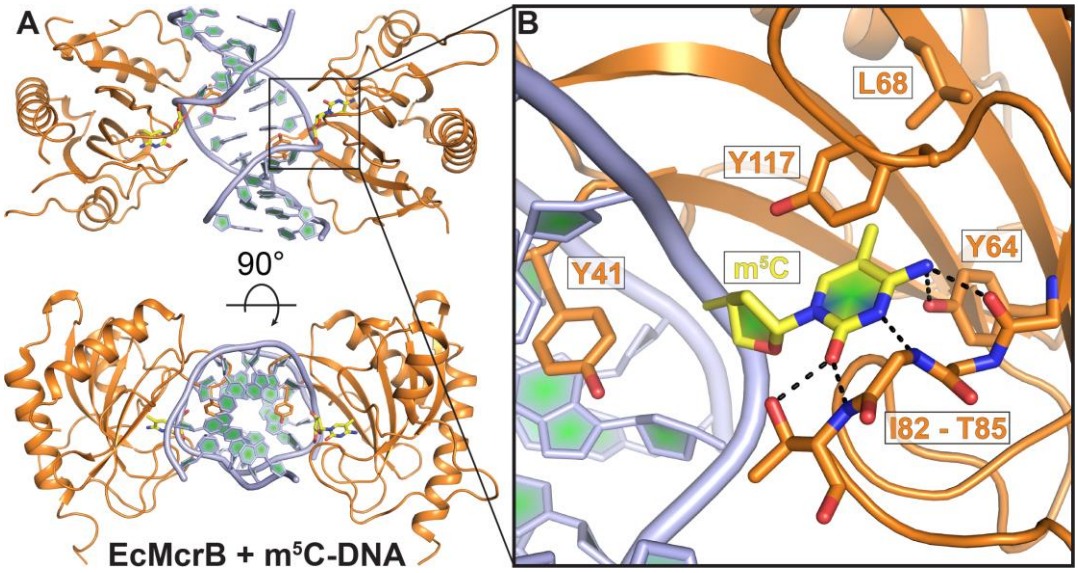

Figure 4. Crystal structure of EcMcrB bound to m^{5}C DNA. *A.* Cartoon representation of EcMcrB (orange) bound to m^{5}C-DNA (light blue) shown in two, perpendicular orientations. *B.* Zoomed in view of the flipped out m^{5}C base stabilized within McrB. The resulting gap is stabilized by insertion of Y41.

achieved through base-flipping in the SRA domain (Han et al., 2015). However, fluorescent studies suggest *Eco*K12 McrA does not bind modified DNA via base-flipping as originally believed (Czapinska et al., 2018).

The McrBC restriction system is the second half of the *E. coli* immigration control system and contains the *mcrB* and *mcrC* genes that encode for McrB and McrC respectively. McrB consists of an N-terminal putative DNA binding domain (McrB-N) and a C-terminal GTP-dependent AAA+ (**A**TPases **a**ssociated with various cellular **a**ctivities) motor domain (McrB AAA+) (Figure 3). A cryptic translational start site in *mcrB* encodes for a shorter version of McrB including only the motor domain and is thought to play a regulatory role in inhibiting McrBC activity (Dila et al., 1990). McrC consists of an N-terminal domain of unknown function (DUF), presumably for GTPase stimulation, and a C-terminal canonical PD-(D/E)XK endonuclease domain. DNA binding occurs through McrB-N and has been shown to recognize DNA containing m^{4}C,

m^5C, or hm^5C in RmC sites (Krüger et al., 1995). A crystal structure of the *E. coli* McrB-N (EcMcrB-N) in complex with m^5C DNA reveals that recognition is achieved via base-flipping of the m^5C out of the DNA duplex (Figure 4A) (Sukackaite et al., 2012). The m^5C base is stabilized within McrB through several polar (Y64, I82 – T85) and hydrophobic/aromatic (L68 and Y117) contacts (Figure 4B). The resulting gap in the DNA duplex is then stabilized by insertion of the Y41 residue (Figure 4B). Mutagenesis of Y41 abolishes EcMcrB-N DNA binding *in vitro* (Sukackaite et al., 2012). Furthermore, phylogenetic analysis of McrB homologs indicate that McrB-N is highly divergent among both bacteria and archaea. Together, these data suggest that homologs of EcMcrB-N may utilize alternate modes of substrate recognition or bind different targets.

Unlike McrB-N, McrB AAA+ is highly conserved. Like other AAA+ proteins, McrB can also undergo nucleotide specific oligomerization. Estimated molecular weights of both McrB or McrB AAA+ in complex with GTPγS by SEC suggests it oligomerizes as a heptamer (Panne et al., 2001). Preliminary scanning transmission electron microscopy (STEM) also supports this stoichiometry (Panne et al., 2001). McrC can only interact with oligomeric McrB with a stoichiometry of 5:1 or 7:2 of B:C as evidenced by SEC and STEM (Panne et al, 2001). Furthermore, previous binding and kinetic studies show that, 1) McrB binds GTP with an affinity of 10^{-6} M^{-1}, 2) McrB binds GDP >50-fold and ATP >1000-fold weaker than GTP, 3) McrB hydrolyzes GTP with a steady-state rate of ~0.5 min^{-1}, and 4) McrC stimulates GTP hydrolysis ~30-fold independent of DNA (Pieper et al., 1997). In fact, no difference in GTP binding or kinetics have been observed between McrB and McrB AAA+ suggesting that GTP

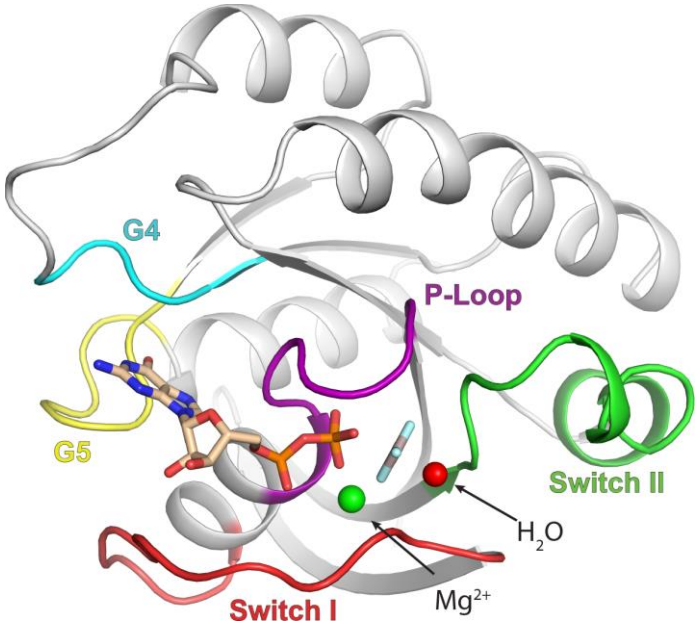

Figure 5. The five conserved GTPase motifs of Ras. Cartoon representation of the small GTPase Ras. The five key GTPase motifs, P-Loop (purple), Switch I (red), Switch II (green), G4 (cyan), and G5 (yellow) are colored accordingly. The bound GDP-AlF_3 is shown in sticks and colored wheat. The Mg^{2+} cofactor and catalytic water are shown in spheres.

hydrolysis is mutually exclusive from DNA binding. Unfortunately, the structural and biochemical underpinnings of GTP binding, hydrolysis, and GTPase stimulation in McrBC remains elusive. Future studies are required to tease apart how this unique AAA+ domain can use GTP as its substrate.

Previous genetic characterization of the *mcrBC* locus revealed highly conserved sequences within McrB AAA+ predicted to be the GTP-binding motif: GxxxxGK (position 201-208), DxxG (position 300-303), and NxxD (333-336) (Dila et al., 1990). To better understand the subtleties underlying the cryptic nature of the McrBs AAA+ GTP specificity, a structural comparison between ATPases and GTPases must first be drawn. GTPases contain five key motifs, G1-G5, and are all involved in nucleotide binding. Ras is a canonical example and used to illustrate the five motifs (Figure 5) (Scheffzek et al., 1997). The G1 motif forms the P-loop with the consensus sequence

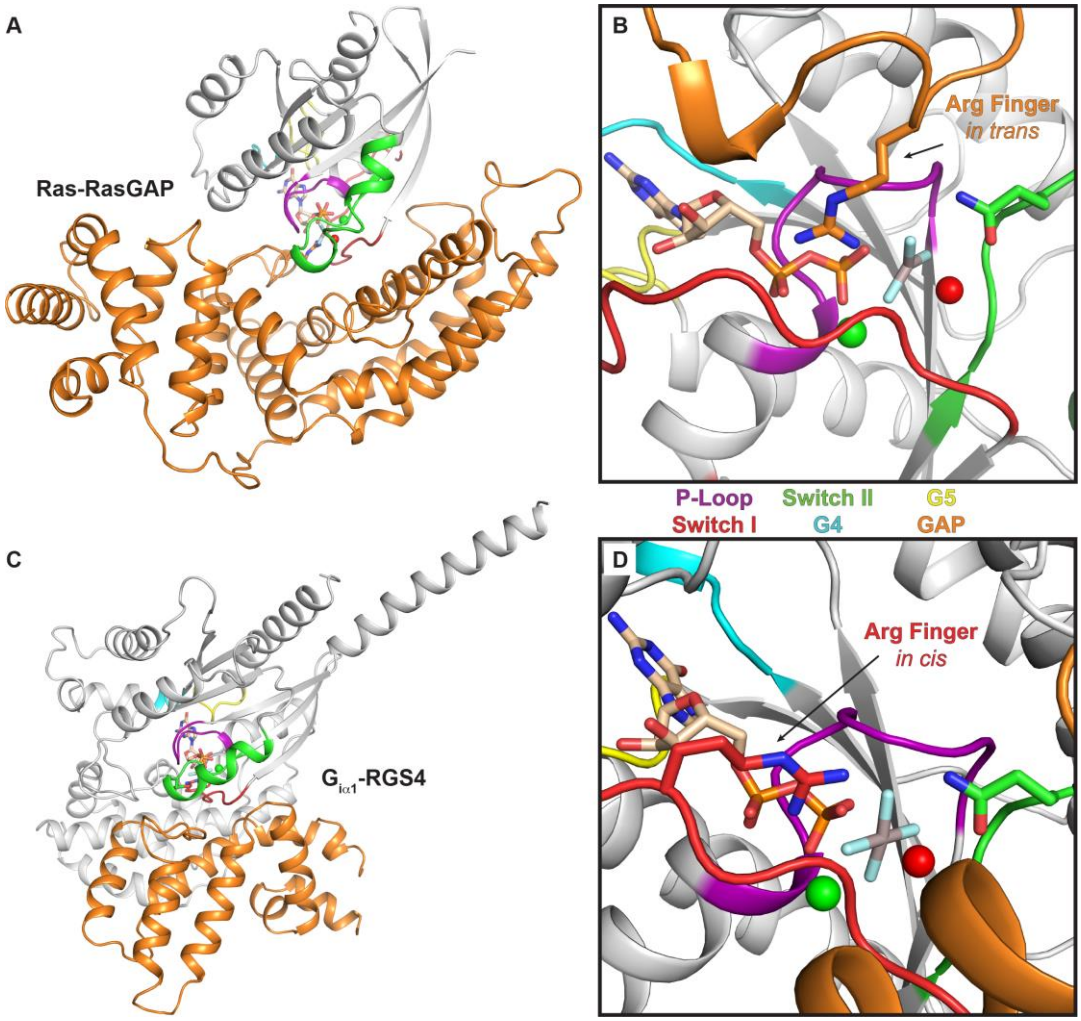

Figure 6. Structural comparison of the Ras-RasGAP and $G_{i\alpha1}$-RGS4 complexes. *A*. Cartoon representation of Ras (white) in complex to its GAP, RasGAP (orange). *B.* Zoomed in view of the active site of Ras-RasGAP. The arrow points to the catalytic Arg Finger (orange) from RasGAP oriented *in trans*. *C.* Cartoon representation of $G_{i\alpha1}$ (white) in complex to its GAP, RGS4 (orange). *D.* Zoomed in view of the active site of $G_{i\alpha1}$-RGS4. The arrow points to the catalytic Arg Finger (red) from $G_{i\alpha1}$ oriented *in cis*. All key structural elements are color coordinated and labeled accordingly.

GxxxxGKS/T. A conserved Thr in G2 and a direct or water-mediated contact by a conserved Asp in G3 participate in Mg^{2+} binding. An additional conserved Gln in G3 positions the catalytic water. Both G2 and G3 undergo γ-phosphate-dependent conformational changes and have been renamed switch I and switch II accordingly. The G4 motif contains the consensus N/TKXD where the Asp mediates specific binding to

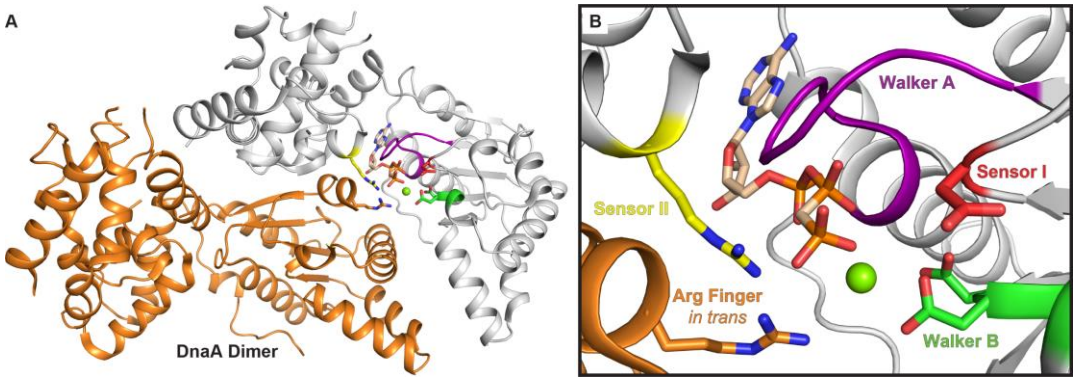

Figure 7. Crystal structure of the DnaA dimer. *A.* Cartoon representation of the DnaA dimer. The subunit bound to AMPPCP is colored white and the subunit providing the Arg finger *in trans* is colored orange. *B.* Zoomed in view of the active site of the DnaA dimer. The five key ATPase motifs, Walker A (purple), Sensor I (red), Walker B (green), Sensor II (yellow), and Arg finger (orange) are colored accordingly.

the guanine base. Mutation of this Asp to Asn confers specificity to XTP. The G5 motif is not entirely well conserved and works to provide additional stabilizing interactions to the guanine base and/or ribose (Daumke et al., 2016).

Two additional proteins are generally required to regulate the GTPase cycle: 1) a GTPase activating protein (GAP) to stimulate GTPase activity and 2) a guanine nucleotide exchange factor (GEF) to facilitate exchange of GDP to GTP. A catalytic Arg residue (Arg finger) is also required for stimulated GTP hydrolysis and stabilizes the charge in the transition state. The crystal structures of many G-proteins coupled to their respective GAPs have been solved and define several canonical mechanisms for GTPase stimulation. In Ras-RasGAP, the Arg finger resides within the GAP and its association to the GTPase drives stimulation (Figure 6A-B) (Scheffzek et al., 1997). Alternatively, in $G_{i\alpha1}$-RGS4, the Arg finger is intrinsic to $G_{i\alpha1}$ and association to its GAP correctly positions the switch regions to drive stimulation (Figure 6C-D) (Tesmer et al., 1997). Other modes of stimulation have also been observed independent of a GAP as found in the dynamin family of GTPases which achieve GTPase stimulation by

homodimerization (Chappie et al., 2010).

AAA+ proteins stem from the ASCE (**a**dditional **s**trand **c**onserved **E**) family of proteins as part of the C-terminal bundle classification (as opposed to the RecA-like ATPases from the C-terminal *hairpin* classification). Within the AAA+ superfamily is a subcategory of different "clades", defined by different insertions of secondary structural elements within the core fold. For example, McrB resides within the H2-insert clade which is defined by a unique β-hairpin insertion in helix 2. Like their GTPase counterparts, AAA+ proteins also contain several key structural motifs (Erzberger et al., 2006). DnaA is a canonical example and used to illustrate these motifs (Figure 7A-B) (Erzberger et al., 2006). Walker A forms the P-loop with the consensus sequence GxPGxxKS/T (G1). Walker B contains a conserved DE motif to coordinate Mg^{2+} binding (switch I). Sensor I contains a conserved Asn used to position the catalytic water (switch II). Sensor II contains a conserved GAD and is involved in adenine base/ribose binding (G5) (Erzberger et al., 2006). Sensor II often contains an Arg that form stabilizing interactions with the α and β phosphates as seen in DnaA (Figure 7B) (Erzberger et al., 2006). AAA+ proteins also utilize an Arg finger for charge compensation that is usually provided via dimerization with another AAA+ protomer; dimerization forms composite active sites as illustrated in DnaA (Figure 7B). Organization of AAA+ oligomers typically form hexameric rings as seen with the minichromosome maintenance protein complex (MCM helicase) or helical filaments as seen with DnaA.

Structurally, McrB is proposed to be a member of the AAA+ family. However, two of the three GTP-binding motifs initially proposed by Dila *et al*. were validated by

mutational analysis (Pieper et al., 1997 and 2002). The third motif, NTAD (NxxD, position 333-336), replaces sensor I in the AAA+ fold. The N333A mutant has no detectable DNA cleavage activity and severely diminished GTP binding and hydrolysis compared to the wild-type protein (Pieper et al., 1997). Homology of NTAD to the canonical G4 element (N/TKxD) of GTPases also led investigators to study the ability of D to N mutants to bind GTP. Interestingly, D336N mutants did not exhibit any DNA cleavage activity with GTP or XTP, and in fact, did not confer XTP specificity as seen in other GTPases. Furthermore, D336N mutants could still hydrolyze GTP despite being unable to be stimulated by McrC (~30% of wild-type activity) and explaining its inability to cleave DNA (Pieper et al., 1999). These results were almost paradoxical and left researchers questioning whether the conserved NTAD sequence was involved in guanine base binding or GTP hydrolysis, two independent events that usually require two motifs (G4 and switch I respectively) to accomplish.

The GAP-like ability of McrC to stimulate McrBs GTPase activity has been less extensively studied. Two possible mechanisms of stimulation have been proposed: 1) McrC contains a key catalytic residue that upon binding the McrB oligomer, is positioned *in trans* (like Ras-RasGAP) or 2) McrC binding stabilizes flexible 'switch' regions in McrB orienting the catalytic machinery *in cis* (like G_{ia1}-RGS4). Few attempts have been made to probe this hypothesis with mutational analysis, and those that have been attempted remain inconclusive. Unfortunately, in the absence of an atomic resolution structure, the biochemical means of GTP binding, hydrolysis, and GTPase stimulation in McrBC remains elusive.

The ability of McrBC to restrict DNA, however, has been well-characterized

biochemically and strongly resembles the Type I RM systems. A mutational analysis of the PD-(D/E)XK motif suggests D244, D257, and K259 from the catalytic center of *E. coli* McrC (Pieper et al., 2002). Successful double-strand cleavage by EcMcrBC requires that two RmC sites (Sutherland et al., 1992; Krüger et al., 1995) be separated by 30-3000 bp (Sutherland et al., 1992; Stewart et al., 1998; Pieper et al., 2002). These sites can exist on either strand (Sutherland et al., 1992; Stewart et al., 2000) with an ideal cleavage length of ~40-60 bp (Pieper et al., 2002). They may also exist on different daughter strands across a replication fork (Ishikawa et al., 2011). Cleavage only occurs ~30-35 bp from one of the two modified bases unless a translocation block is encountered (Panne et al., 1999). Long range (>80 bp) cleavage requires GTPase hydrolysis and stimulation (Stewart et al., 1998). Monomeric or oligomeric McrB can associate with each site independent of McrC

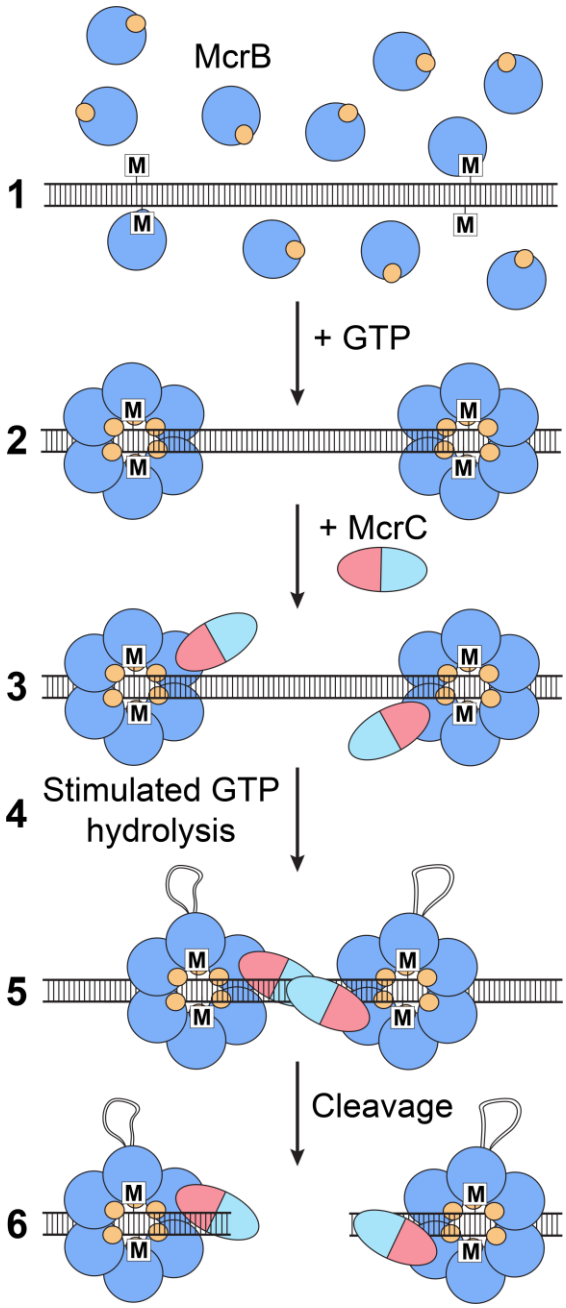

Figure 8. Proposed model for McrBC GTP-stimulated hydrolysis and cleavage. The six steps involved in successful cleavage are numbered accordingly.

(Panne et al., 2001; Pieper et al., 2002).

Over the past 3 decades, the combined efforts of many investigators attempting to understand the GTP-stimulated hydrolysis and cleavage by McrBC has yielded a working model for DNA restriction (Figure 8):

1. McrB binds two RmC sites separated by 30-3000 bp
2. McrB undergoes GTP-dependent oligomerization
3. The McrB oligomer recruits McrC
4. McrC stimulates McrBs GTPase activity
5. Stimulation of GTP hydrolysis leads to DNA translocation
6. Collision of two assemblies near one RmC site results in cleavage

Although much of this model has been extensively supported by biochemistry, many questions remain:

1. How do homologs of McrB bind DNA?
2. How does McrB bind GTP?
3. How does McrC bind and stimulate GTP hydrolysis?
4. How does translocation occur?

Further efforts must be taken to fully address these questions.

One major complication in characterizing McrBC is its dependence on DNA modification. An McrBC related homolog, *Lla*JI, has been identified on pNP40, a naturally occurring 65 kb plasmid from *Lactococcus lactis* (*L. lactis*) (O'Driscoll et al., 2004). The operon encoding *Lla*JI contains four proteins: two m^5C methyltransferases, M1 and M2, and two restriction proteins, R1 and R2 (Figure 9A). *Lla*JI.R1 consists of an N-terminal putative DNA binding domain (*Lla*JI.R1 Δ226) and a C-terminal AAA+

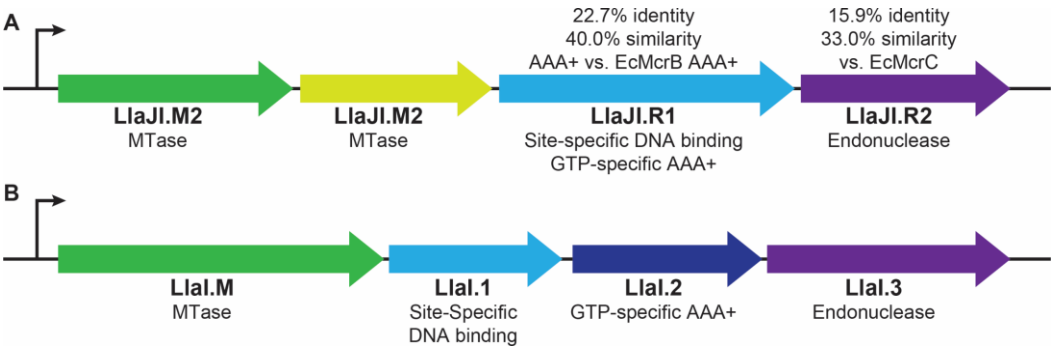

Figure 9. Organization of the *Lactococcus lactis Lla*JI and *Lla*I operons. *A.* Organization of the *Lla*JI operon on the pNP40 plasmid in *L. lactis*. The operon encodes for four proteins: two m^5C MTases, M1 and M2, and two Reases, R1 and R2. Percent identity and similarity of *Lla*JI.R1 and R2 to McrB and C are shown. *B.* Organization of the *Lla*I operon on the pTR2030 plasmid in *L. lactis*. The operon encodes for four proteins: one m^5C MTase and three restriction proteins, 1-3.

(*Lla*JI.R1 226-585). *Lla*JI.R2 consists of an N-terminal DUF and a C-terminal PD-(D/E)XK endonuclease. The McrB signature NTAD sequence is also conserved in *Lla*JI.R1 and suggests that *Lla*JI.R1 and R2 are McrB and C homologs respectively. The two methyltransferases, M1 and M2, have been shown to methylate the asymmetric 5'-GACGC-3' and complementary 5'-GCGTC-3' respectively (O'Driscoll et al., 2005). A homolog of *Lla*JI in *Clostridium cellulovorans* (*C.cellulovorans*), *Cce*743, has also been identified with the same specificity (Yang et al., 2016). Both sequences exist within the *Lla*JI.R1 and R2 promoters and are transcriptionally regulated by methylation (O'Driscoll et al., 2005). Reconstitution of *Lla*JI.R1 *in vitro* confirms it recognizes the same asymmetric 5'-GACGC-3' and 5'-GCGTC-3' site-specifically as M1 and M2 and is blocked by methylation (O'Driscoll et al., 2006). Additionally, both *Lla*JI.R1 and R2 are required for cleavage *in vivo*. The *Lla*JI restriction cassette therefore resembles classical Type IIS RM systems but may behave in a similar fashion as the Type IV MDR, McrBC.

Another related RM system, *Lla*I, was identified on pTR2030, a 46.2 kb

conjugative plasmid from *L. lactis.* The operon encoding *Lla*I also contains four proteins: one methyltransferase, M, and three restriction proteins, 1-3 (Figure 9B) (O'Sullivan et al., 1995). Unlike McrBC and *Lla*JI, *Lla*I contains three restriction proteins: *Lla*I.1 is predicted to be a putative DNA binding domain, *Lla*I.2 is a AAA+ (with conserved NTAD), and *Lla*I.3 is a two domain protein with an N-terminal DUF and C-terminal PD-(D/E)XK endonuclease. *In vivo* studies show that *Lla*I.1 is essential for restriction while *Lla*I.2 and *Lla*I.3 allowed for inefficient restriction of phage DNA to occur (re O'Sullivan et al., 1995). A frameshift mutation in *Lla*I.M proved lethal to *L. lactis* implying that restriction is active without the M subunit and that methylation by *Lla*I.M protects the host from self-cleavage (O'Sullivan et al., 1995). A homolog of *Lla*I in *Bacillus subtilis* (*B.subtilis*), *Bsu*MI, has also been identified as an isoschizomer to *Xho*I (CTCGAG) (Jentsch et al.. 1983). The operon encoding BsuMI resembles that of *Lla*I and includes four proteins with similar predicted functions. Together, the *Lla*JI, *Lla*I, and *Bsu*MI systems provide a platform for studying McrBC-related mechanisms in a modification-independent manner.

The McrBC restriction system describes only one of the few MDR systems that have been identified. In fact, of these, very few have been biochemically characterized with even fewer high-resolution structures reported (including of the enigmatic Type IIM systems). Mechanisms of DNA methylation-dependent cleavage are therefore not very well understood as whole. In the past decade, several crystal structures have emerged that yield significant insight into both modification-specific DNA recognition and cleavage. Of these are the Type IIM RM systems, *Dpn*I and MspJI, and the Type IV MDR system, PvuRts1I. Although crystal structures are not yet available for the

Type IV MDR systems, Mrr and GmrSD, significant biochemical characterization on them has been made on and will also be discussed.

The Mrr protein was first described in 1987 when expression of the site-specific adenine MTases, *Hha*II (Gm^6ANTC) or *Pst*I ($CTGCm^6AG$), induced the RecA-dependent SOS DNA repair response in *E. coli* by forming double-strand breaks (DSB) (Heitman et al., 1987). This result is similar to the expression of site-specific cytosine MTases in *E. coli* inducing the SOS response as a consequence of McrBC activity, suggesting that Mrr is an endonuclease that targets DNA containing m^6A. Mrr was also discovered to be activated upon heterologous expression of foreign MTases, such as HhaII (Tesfazgi Mebrhatu M et al., 2011). Interestingly, a recent study showed that a sub-lethal hydrostatic pressure (HP) shock of ~100 MPa also elicits the SOS response. *In vivo* fluorescence studies of GFP-Mrr not only revealed that 1) HP shock triggers Mrr activity by forcing inactive Mrr tetramers to dissociate into active dimers, but also that 2) the HhaII MTase triggers Mrr activity by creating high affinity target sites on the chromosome, pulling the equilibrium from tetrameric to dimeric (Bourges et al., 2017). Together these data reveal a control mechanism for selective activation of Mrr driven through its oligomerization. To date, an atomic resolution structure of Mrr has not been solved, however, recent bioinformatic analyses supported by mutagenesis showed that it is composed of a winged helix (wH) and PD-(D/E)XK domains.

The Type IIM REase, *Dpn*I, is widely used as a tool in molecular biology for its ability to cleave modified DNA in a site-specific manner. It targets and cleaves the short, palindromic sequence $G(m^6A)TC$ and is 5-fold more efficient at cleaving fully-methylated DNA than hemi-methylated (Siwek et al., 2012). The crystal structure of

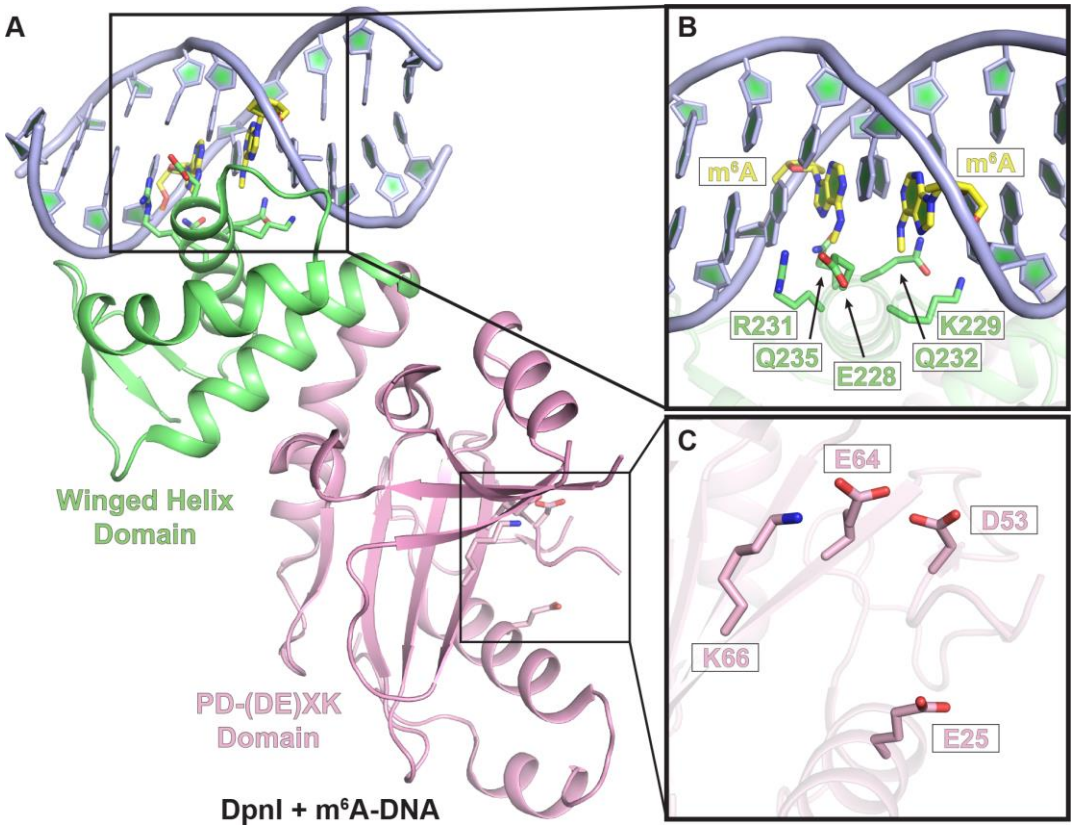

Figure 10. Crystal structure of *Dpn*I bound to m^{6}A DNA. *A.* Cartoon representation of *Dpn*I bound to m^{6}A DNA. The winged helix domain (green), PD-(D/E)XK domain (pink), and m^{6}A DNA (light blue) are colored accordingly. *B.* Zoomed in view of the DNA binding site. The m^{6}A bases are colored yellow and shown in sticks. Residues important for DNA binding are colored green and shown in sticks. *C*. Zoomed in view of the PD-(D/E)XK active site. Residues important for DNA cleavage are colored pink and shown in sticks.

*Dpn*I bound to its cognate substrate was solved to 2.05 Å and revealed that *Dpn*I, like Mrr, utilizes a winged helix (wH) domain to bind DNA (Figure 10A, green) (Siwek et al., 2012). The C-terminal, 'recognition helix', inserts directly into the major groove where the two m^{6}A moieties are closest together (Figure 10B). An extensive network of hydrogen bonds is formed between the residues of the recognition helix and the purine bases within the recognition site. Unlike most RM or MDR systems that oligomerize to form a cleavage competent complex, *Dpn*I has been shown to cleave DNA as a monomer. Interestingly, the conserved PD-(D/E)XK domain (Figure 10C) is located far

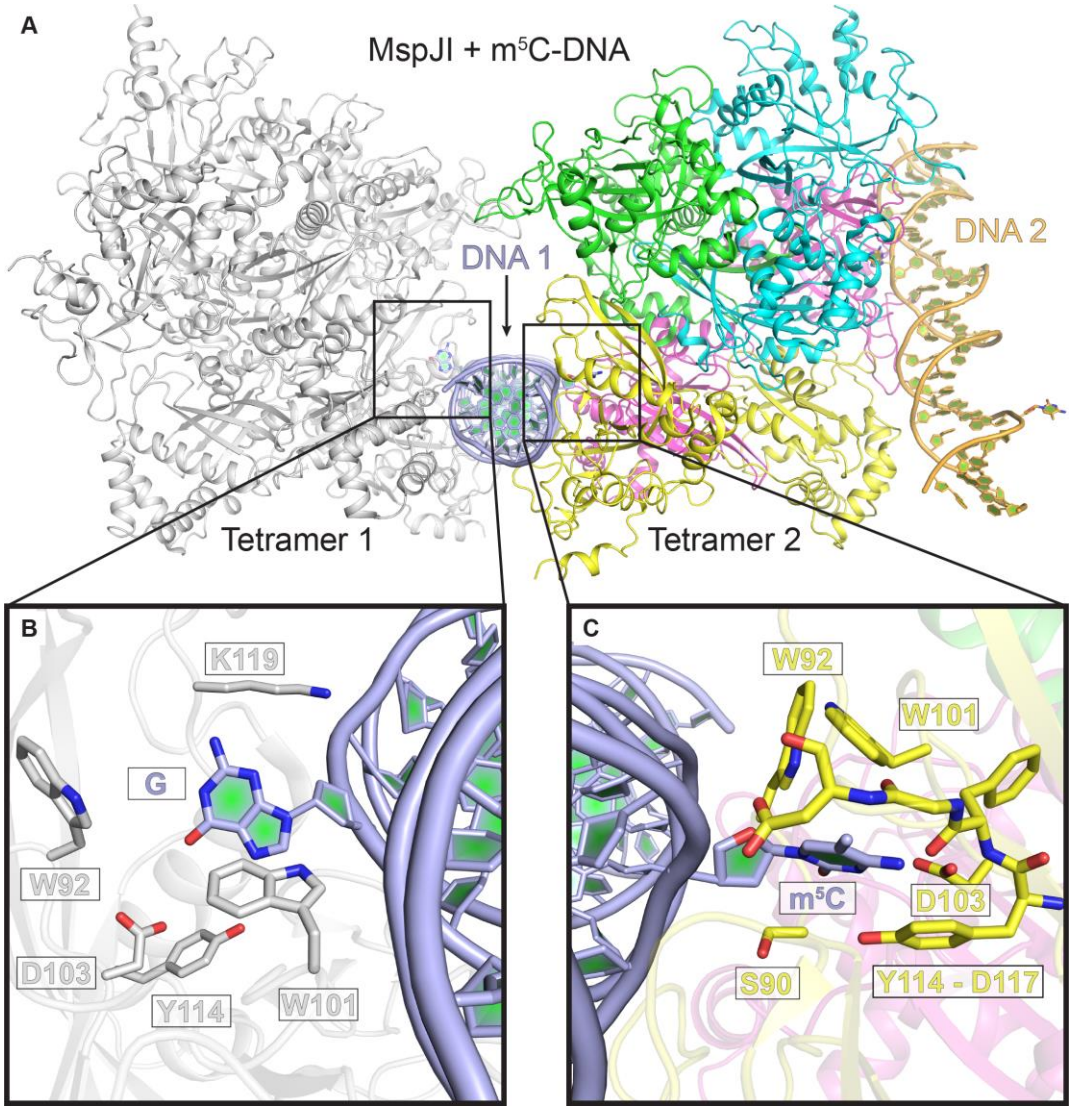

Figure 11. Crystal structure of MspJI bound to m^{5}C DNA. *A.* Cartoon representation of MsPJI bound to m^{5}C DNA. Both tetramers are shown bound to DNA molecule 1 (light blue) while only tetramer 2 is shown bound to both DNA molecules 1 and 2. (light blue and orange respectively). The four chains in tetramer 2 are colored differently for clarity. Only the yellow and pink subunits interact with DNA. *B.* Zoomed in view of the m^{5}C binding pocket nonspecifically bound to a G base. Residues forming stabilizing contacts are shown in sticks and colored grey. *C*. Zoomed in view of the m^{5}C binding pocket bound to m^{5}C. Residues forming stabilizing contacts are shown in sticks and colored yellow.

from the recognition site and suggests that *Dpn*I must either undergo drastic conformational changes or cleave at a different site. Indeed, structural modeling shows that *Dpn*I likely binds at one site and cleaves at another (Siwek et al., 2012).

Like *Dpn*I, MsPJI also belongs to the Type IIM family of REases and cleaves

modified DNA. It recognizes the m^5CNNR or hm^5CNNR (where N = any nucleotide and R = purine, A or G) and cleaves downstream ~9 bases away on the modified strand and ~13 bases away on the complementary strand, leaving a ~9/13 cleavage product with short 3'-overhangs (Cohen-Karni et al., 2011; Zheng et al., 2010). MspJI is reminiscent of the Type IIS family of REases which have a similar domain architecture and cleave outside of their recognition sequence. These REases often behave cooperatively and are catalyzed by two or more allosterically regulated molecules, typically due to the independently acting active sites. MspJI, for instance, is stimulated by the addition of oligonucleotides that contain its recognition sequence (Cohen-Karni et al., 2011; Zheng et al., 2010). A crystal structure of the MspJI apoprotein reveals that it forms a homotetramer where two molecules are close together ('closed' conformation) while two are farther apart ('open' conformation) and supports the idea of allostericity (Horton et al., 2012). This is further corroborated by the crystal structure of the m^5C DNA bound complex where the homotetramer is shown to interact with two DNA molecules while each DNA molecule is bound to two homotetramers (Figure 11A) (Horton et al., 2014). Interestingly, only two molecules (yellow and pink) of the homotetramer ever engage the DNA in the 'open' conformation (Figure 11A). Each of these two molecules therefore forms two interactions, one with each of the two bound DNA duplexes. For instance, the yellow molecule directly interacts with the m^5C base via its N-terminal domain on one DNA duplex while poising its C-terminal PD-(D/E)XK domain towards the other duplex (Figure 11C). The other molecule, however, nonspecifically flips out the base 13 bp downstream in the complementary strand (Figure 11B). This base is only partially stabilized within the m^5C binding pocket and

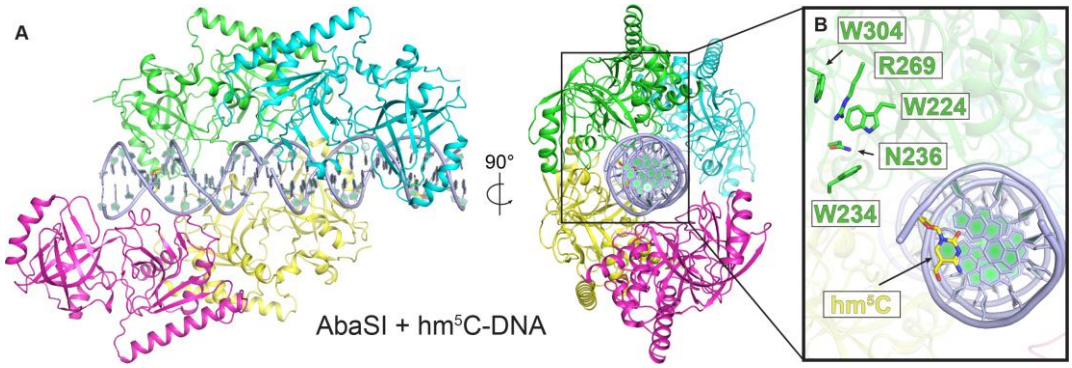

Figure 12. Crystal structure of *Aba*SI bound to hm^{5}C DNA. *A.* Cartoon representation of *Aba*SI bound to hm^{5}C DNA shown in two, perpendicular orientations. Two dimers are shown bound to DNA (Green+Blue and Magenta+Yellow) The bound DNA is colored light blue. *B.* Zoomed in view of the hm^{5}C (yellow) found intra-helically. The conserved hm^{5}C binding cavity in the SRA domain is shown in sticks and colored green.

forms different interactions than the bound m^{5}C (Figures 11B-C). Base-flipping of the m^{5}C (and nonspecific) base is accomplished via an SRA (**S**ET and **R**ING-**a**ssociated) domain which has been shown to bind m^{5}C DNA via base-flipping previously (Hashimoto et al., 2008). The non-specific binding and concomitant base-flipping by MspJI was therefore unexpected and shrugged off as part of the sequence-discrimination mechanism.

The PvuRts1I-family of MDR systems cleaves DNA containing hm^{5}C and glucosylated hm^{5}C (ghm^{5}C). *Aba*SI is a prototypical member and has been extensively characterized. It recognizes ghm^{5}C far more efficiently than hm^{5}C or nm DNA by selectivity factors of 8000:500:1 respectively (Wang et al., 2011). Like MspJI, *Aba*SI binds its site and cleaves with some variability downstream ~9 bases away on the modified strand and ~13 bases away on the complementary strand, leaving a ~9/13 cleavage product with short 3'-overhangs. Optimal cleavage occurs when two ghm^{5}C bases occur 21-23 bp apart on opposite strands (Borgaro et al., 2013). A crystal structure of *Aba*SI in the absence of DNA reveals that it dimerizes through its N-terminal Vsr-

like endonuclease domain and that the C-terminus adopts an SRA domain (Horton et al., 2014). Dimerization spatially orients the hm^5C binding pockets within the SRA domains by ~70 Å which is consistent with the ~22 bp separation of its binding sites. A complementary crystal structure in the presence of hm^5C DNA shows two dimers in the asymmetric unit nonspecifically interacting with DNA with the modified base remaining intra-helical (Figure 12A) (Horton et al., 2014). Interestingly, both the endonuclease and hm^5C binding sites are localized far from the DNA and suggest significant conformational changes or DNA bending are required for a cleavage competent complex to form (Figure 12B).

Another Type IV MDR, GmrSD, has been identified to target and cleave hm^5C and ghm^5C DNA. GmrSD is comprised of two subunits, GmrS and GmrD, encoded by the *gmrS and gmrD* genes respectively (Bair et al., 2007). In most bacteria, however, the *gmrS* and *gmrD* genes are fused together and encode for a single-chain protein. In the absence of an atomic resolution structure, bioinformatic and mutational studies suggest that GmrSD contain two conserved protein domains, an N-terminal DUF and a proposed C-terminal HNH nuclease. Interestingly, while no NTPase is predicted within GmrSD, its cleavage activity has been shown to be stimulated by the presence of ATP/GTP (He et al., 2015). Although little is known about GmrSD function, the T4 phage encoded protein, Internal Protein I* (IPI*), has been demonstrated to bind and inhibit GmrSD (Bair et al., 2007). Additional phage encoded inhibitors of MDR systems have been identified, including the anti-restriction nuclease (Arn), that can inhibit McrBC function (Dharmalingam et al., 1982). The structure of Arn reveals it is a DNA mimetic that potentially disrupts McrBC function by precluding DNA binding (Ho et

al., 2014). Analysis of these MDR systems and their inhibitors reveals an evolutionary pathway that has elaborated a diverse and specific set of coevolving attack and defense structures. They are of extreme interest to researchers in the field as they provide a weakness that can potentially be exploited for therapeutics to improve the efficacy of phage therapy.

PHAGE EXCLUSION SYSTEMS

All the systems described up to this point, including RM and MDR systems, are colloquially defined as 'restriction' systems. The term 'exclusion', however, is a more general term that describes the inability of a phage to successfully infect bacteria, whereas 'restriction' is used to describe specific nuclease digestion of invading DNAs. It can therefore be stated that all restriction systems are exclusion systems, but not all exclusion systems are restriction systems. Several such exclusion systems have been previously characterized and include Rex, PrrC, Lit and BREX. Phage exclusion by these systems is achieved by activation of a prophage-encoded toxin that mediates bacterial cell death, thereby preventing viral propagation (Snyder et al., 1995; Gottesman et al., 1998).

The Rex system is comprised of the two proteins, RexA and RexB, encoded by the *rexA* and *rexB* genes of bacteriophage λ respectively. Together they create a two-component system that aborts lytic growth of bacterial viruses. Low levels of RexB are expressed in the lysogenic state and increased 2 to 10-fold after superinfection by a competing phage (Parma et al., 1992). Previous biochemical characterization of RexB suggests it to be a polytopic transmembrane protein that forms ion channels in response

to lytic phage growth, thereby depolarizing the cell and inducing altruistic cell death (Parma et al., 1992). The role of RexA in phage exclusion has been less clear. Overexpression of RexA in the presence of RexB elicits an exclusion-like response in uninfected cells (Snyder et al., 1989). This led researchers to hypothesize RexA would localize to RexB thereby activating its ion channel activities. Recent studies in our lab, however, suggest that RexA may be acting independently of RexB and instead interacts with the cI repressor protein that represses lytic promoters in the λ lysogen. Further studies are being conducted to probe this mechanism.

PrrC is a $tRNA^{Lys}$ anticodon nuclease (ACNase) encoded by the optional *E. coli prr* locus. Its ACNase activity is silenced via association with the REase *Eco*prrI. A phage T4-encoded inhibitor of *Eco*prrI, Stp, activates the latent ACNase activity resulting in cleavage of the $tRNA^{Lys}$ thereby resulting in cell death and phage exclusion (Levitz et al., 1990). Lit is a protease encoded by *e14*, a cryptic prophage element in *Eco*K12. Although Lit is constitutively expressed in *E. coli*, it lies dormant in the absence of its activator, the Gol sequence of the T4 phage capsid protein, gp23 (Yu and Snyder, 1994; Vallee and Auld, 1990). Activation of Lit by gp23 results in cleavage of the TGITI motif of Ef-Tu, inhibiting protein synthesis and blocking viral replication (Bergsland et al., 1990).

The Pgl (**p**hage **g**rowth **l**imitation) system of *Streptomyces coelicolor* is another exclusions system that affects the growth of phage ϕC31. It consists of the four genes *pglWXYZ*, where PglW may be a kinase, PglX is a SAM-dependent MTase, PglY is an ATPase, and PglZ is an alkaline phosphatase (Hoskisson et al., 2015). The current model of this system suggests that PglXZ act as a toxin-antitoxin pair such that release

of PglX in ϕC31-infected cells is controlled by PglZ and leads to formation of active RM complexes (Hoskisson et al., 2015). Recent bioinformatic analysis of prokaryotic defense islands revealed overrepresentation of genes encoding PglZ homologs and has been renamed BREX (for **b**acte**r**iophage **ex**clusion) (Makarova et al., 2011; Goldfarb et al., 2015). Introduction of BREX into *Bacillus cereus* without its own *brx* genes resulted in increases levels of resistance against diverse phages where phage replication was blocked without restriction (Gordeeva et al., 2019). Counterintuitively, the host genomic DNA was also methylated while the phage genome remained unmethylated. The overall mechanism of how BREX excludes phage remains to be seen.

CRISPR-CAS

CRISPR-Cas (**c**lustered **r**egularly **i**nterspaced **s**hort **p**alindromic **r**epeats) systems are prokaryotic, adaptive immune systems originally discovered in 1987 by Yoshizumi Isino and others (Ishino et al., 1987). They function in two major steps:

1. Acquisition of phage sequences – The Cas1-Cas2 proteins target the phage DNA and excise a small fragment, called a protospacer. The protospacer is then integrated within a repeating CRISPR array within the host genome next to a PAM (**p**rotospacer **a**djacent **m**otif) site (Ishino et al., 2018).
2. Immunity against reinfection – Upon reinfection, the CRISPR array is transcribed into short RNA sequences used by the Cas9 protein. Cas9 can then utilize the RNA sequence as a targeting mechanism to the re-infecting DNA for cleavage (Ishino et al., 2018).

CRISPR-Cas varies from canonical RM and MDR systems in that they use

complementary RNA to accomplish specificity instead of protein-DNA interactions. Since its discovery, the CRISPR-Cas system has been heavily engineered as a molecular tool for gene editing.

SIGNIFICANCE

Although the bacterial defense systems discussed above yield significant insight into their mechanisms of action, they represent only a handful of the thousands of systems known. In the context of bacterial restriction systems, the aim to understanding their function is largely 2-fold:

First, evolutionarily, it has been challenging for bacteria to adapt different site-specific targets. This is due to the targeted recognition of the REase and MTase acting at the same site. For instance, if the REase evolved a different specificity, then the MTase would lag and the host would die due to improper methylation (or vice versa). It is therefore not surprising that for a simple six bp recognition site which contains a possible 4^6 (4,096) possible sequences, that only several hundred REases exist with different specificities and that the majority of newly discovered REases are isoschizomers (REases with the same specificity). Concomitantly, MDR systems have been routinely used in analyzing the methylation status of genomic DNA, making it an important diagnostic tool for epigenetic disorders. The strict regulations underlying MDR function renders each system useful for only a few specific applications (e.g. McrBC is limited to modified cytosine restriction of bases separated great than 40-60 bp apart). Their broad use as molecular tools makes them an attractive topic for research and engineering to widen the range of potential applications (i.e. increasing useable site

specificities or modifications).

Second, the increasing rate of exchange of antibacterial resistance genes is steadily growing out of control. The emergence of antibiotic resistant, superbug-related infections is either difficult or impossible to treat. As such, there is a pressing urgency to find alternative approaches to treating bacterial infection. Phage therapy is one possibility; however, the bacterial defense mechanisms must be overcome for an effective therapeutic to be created. Understanding the function of bacterial defense systems can therefore be exploited to create new drugs or inhibitors to render them less effective. The presence of phage encoded inhibitors, such as IPI* against GmrSD or Arn against McrBC, provides insight into possible mechanisms of inhibition of these systems (Bair, 2007; Dharmalingam, 1982; Ho, 2014).

The objective of my work described in this document aims to elucidate the molecular underpinnings of McrBC function. I specifically explore the evolutionary divergence of DNA recognition of McrB homologs by using a combination of structural biology, biochemistry, and molecular biology. To this end, I have solved the crystal structure of the *Helicobacter pylori* LlaJI N-terminal domain (HpΔ136), the *Thermococcus gammatolerans* McrB N-terminal domain (TgΔ185), and the *Staphylothermus marinus* McrB N-terminal domain (Sm3-180). HpΔ136 is shown to adopt a B3 domain used for site-specific recognition (Chapter 2), TgΔ185 has coopted a YTH domain for m^6A DNA binding (Chapter 3), and Sm3-180 has coopted an EVE domain for non-specific DNA binding (Chapter 4). Together, these data suggest that McrBC is inherently a modular restriction system that can adapt unique N-terminal domains to alter its specificity. I have additionally solved the crystal structure of the

LlaI.2 GTP-specific AAA+ protein that reveals the molecular determinants of GTP specificity and hydrolysis in McrB homologs (Appendix 1).

REFERENCES

Bair CL, Black LW. A type IV modification dependent restriction nuclease that targets glucosylated hydroxymethyl cytosine modified DNAs. J Mol Biol. 2007 Feb 23;366(3):768-78. Epub 2006 Nov 21.

Bergsland KJ, Kao C, Yu YT, Gulati R, Snyder L. A site in the T4 bacteriophage major head protein gene that can promote the inhibition of all translation in Escherichia coli. J Mol Biol. 1990 Jun 5;213(3):477-94.

Borgaro JG, Zhu Z. Characterization of the 5-hydroxymethylcytosine-specific DNA restriction endonucleases. Nucleic Acids Res. 2013 Apr;41(7):4198-206. doi: 10.1093/nar/gkt102. Epub 2013 Mar 12.

Bourges AC, Torres Montaguth OE, Ghosh A, Tadesse WM, Declerck N, Aertsen A,Royer CA. High pressure activation of the Mrr restriction endonuclease in Escherichia coli involves tetramer dissociation. Nucleic Acids Res. 2017 May 19;45(9):5323-5332. doi: 10.1093/nar/gkx192.

Callahan SJ, Luyten YA, Gupta YK, Wilson GG, Roberts RJ, Morgan RD, Aggarwal AK. Structure of Type IIL Restriction-Modification Enzyme MmeI in Complex with DNA Has Implications for Engineering New Specificities. PLoS Biol. 2016 Apr 15;14(4):e1002442. doi: 10.1371/journal.pbio.1002442. eCollection 2016 Apr. Erratum in: PLoS Biol. 2016 May;14(5):e1002471.

Chappie JS, Acharya S, Leonard M, Schmid SL, Dyda F. G domain dimerization controls dynamin's assembly-stimulated GTPase activity. Nature. 2010 May

27;465(7297):435-40. doi: 10.1038/nature09032. Epub 2010 Apr 28.

Cohen-Karni D, Xu D, Apone L, Fomenkov A, Sun Z, Davis PJ, Kinney SR, Yamada-Mabuchi M, Xu SY, Davis T, Pradhan S, Roberts RJ, Zheng Y. The MspJI family of modification-dependent restriction endonucleases for epigenetic studies. Proc Natl Acad Sci U S A. 2011 Jul 5;108(27):11040-5. doi: 10.1073/pnas.1018448108. Epub 2011 Jun 20.

Czapinska H, Kowalska M, Zagorskaite E, Manakova E, Slyvka A, Xu SY, Siksnys V, Sasnauskas G, Bochtler M. Activity and structure of EcoKMcrA. Nucleic Acids Res. 2018 Oct 12;46(18):9829-9841. doi: 10.1093/nar/gky731.

Daumke O, Praefcke GJ. Invited review: Mechanisms of GTP hydrolysis and conformational transitions in the dynamin superfamily. Biopolymers. 2016 Aug;105(8):580-93. doi: 10.1002/bip.22855. Review. Erratum in: Biopolymers. 2018 Feb;109 (2):.

Dharmalingam K, Revel HR, Goldberg EB. Physical mapping and cloning of bacteriophage T4 anti-restriction endonuclease gene. J Bacteriol. 1982 Feb;149(2):694-9.

Dila D, Sutherland E, Moran L, Slatko B, Raleigh EA. Genetic and sequence organization of the mcrBC locus of Escherichia coli K-12. J Bacteriol. 1990 Sep;172(9):4888-900.

Erzberger JP, Berger JM. Evolutionary relationships and structural mechanisms of AAA+ proteins. Annu Rev Biophys Biomol Struct. 2006;35:93-114.
Erzberger JP, Mott ML, Berger JM. Structural basis for ATP-dependent DnaA assembly and replication-origin remodeling. Nat Struct Mol Biol. 2006 Aug;13(8):676-83. Epub 2006 Jul 9.

Goldfarb T, Sberro H, Weinstock E, Cohen O, Doron S, Charpak-Amikam Y, Afik S, Ofir G, Sorek R. BREX is a novel phage resistance system widespread in microbial genomes. EMBO J. 2015 Jan 13;34(2):169-83.

Gordeeva J, Morozova N, Sierro N, Isaev A, Sinkunas T, Tsvetkova K, Matlashov M, Truncaite L, Morgan RD, Ivanov NV, Siksnys V, Zeng L, Severinov K. BREX system of Escherichia coli distinguishes self from non-self by methylation of a specific DNA site. Nucleic Acids Res. 2019 Jan 10;47(1):253-265.

Gottesman S. Protecting the neighborhood: extreme measures. Proc Natl Acad Sci U S A. 1998 Mar 17;95(6):2731-2.

Han T, Yamada-Mabuchi M, Zhao G, Li L, Liu G, Ou HY, Deng Z, Zheng Y, He X. Recognition and cleavage of 5-methylcytosine DNA by bacterial SRA-HNH proteins. Nucleic Acids Res. 2015 Jan;43(2):1147-59.

Handy DE, Castro R, Loscalzo J. Epigenetic modifications: basic mechanisms and role in cardiovascular disease. Circulation. 2011 May 17;123(19):2145-56.

Hashimoto H, Horton JR, Zhang X, Bostick M, Jacobsen SE, Cheng X. The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. Nature. 2008 Oct 9;455(7214):826-9.

He X, Hull V, Thomas JA, Fu X, Gidwani S, Gupta YK, Black LW, Xu SY. Expression and purification of a single-chain Type IV restriction enzyme Eco94GmrSD and determination of its substrate preference. Sci Rep. 2015 May 19;5:9747.

Heitman J, Model P. Site-specific methylases induce the SOS DNA repair response in

Escherichia coli. J Bacteriol. 1987 Jul;169(7):3243-50.
Ho CH, Wang HC, Ko TP, Chang YC, Wang AH. The T4 phage DNA mimic protein Arn inhibits the DNA binding activity of the bacterial histone-like protein H-NS. J Biol Chem. 2014 Sep 26;289(39):27046-54.

Horton JR, Mabuchi MY, Cohen-Karni D, Zhang X, Griggs RM, Samaranayake M, Roberts RJ, Zheng Y, Cheng X. Structure and cleavage activity of the tetrameric MspJI DNA modification-dependent restriction endonuclease. Nucleic Acids Res. 2012 Oct;40(19):9763-73.

Horton JR, Borgaro JG, Griggs RM, Quimby A, Guan S, Zhang X, Wilson GG, Zheng Y, Zhu Z, Cheng X. Structure of 5-hydroxymethylcytosine-specific restriction enzyme, AbaSI, in complex with DNA. Nucleic Acids Res. 2014 Jul;42(12):7947-59.

Horton JR, Wang H, Mabuchi MY, Zhang X, Roberts RJ, Zheng Y, Wilson GG, Cheng X. Modification-dependent restriction endonuclease, MspJI, flips 5-methylcytosine out of the DNA helix. Nucleic Acids Res. 2014 Oct 29;42(19):12092-101.

Hoskisson PA, Sumby P, Smith MCM. The phage growth limitation system in Streptomyces coelicolor A(3)2 is a toxin/antitoxin system, comprising enzymes with DNA methyltransferase, protein kinase and ATPase activity. Virology. 2015 Mar;477:100-109.

Ishikawa K, Handa N, Sears L, Raleigh EA, Kobayashi I. Cleavage of a model DNA replication fork by a methyl-specific endonuclease. Nucleic Acids Res. 2011 Jul;39(13):5489-98.

Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. J Bacteriol. 1987 Dec;169(12):5429-33.

Ishino Y, Krupovic M, Forterre P. History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology. J Bacteriol. 2018 Mar 12;200(7).

Iyer LM, Leipe DD, Koonin EV, Aravind L. Evolutionary history and higher order classification of AAA+ ATPases. J Struct Biol. 2004 Apr-May;146(1-2):11-31.

Jentsch S. Restriction and modification in Bacillus subtilis: sequence specificities of restriction/modification systems BsuM, BsuE, and BsuF. J Bacteriol. 1983 Nov;156(2):800-8.

Keen EC. A century of phage research: bacteriophages and the shaping of modern biology. Bioessays. 2015 Jan;37(1):6-9.

Krüger T, Wild C, Noyer-Weidner M. McrB: a prokaryotic protein specifically recognizing DNA containing modified cytosine residues. EMBO J. 1995 Jun 1;14(11):2661-9.

Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nat Rev Microbiol. 2010 May;8(5):317-27.

Levitz R, Chapman D, Amitsur M, Green R, Snyder L, Kaufmann G. The optional E. coli prr locus encodes a latent form of phage T4-induced anticodon nuclease. EMBO J. 1990 May;9(5):1383-9.

Loenen WA, Dryden DT, Raleigh EA, Wilson GG. Type I restriction enzymes and their relatives. Nucleic Acids Res. 2014 Jan;42(1):20-44.

Luria SE, Human ML. A nonhereditary, host-induced variation of bacterial viruses. J Bacteriol. 1952 Oct;64(4):557-69.

Makarova KS, Wolf YI, Snir S, Koonin EV. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. J Bacteriol. 2011 Nov;193(21):6039-56.

Meisel A, Bickle TA, Krüger DH, Schroeder C. Type III restriction enzymes need two inversely oriented recognition sites for DNA cleavage. Nature. 1992 Jan 30;355(6359):467-9.

McClarin JA, Frederick CA, Wang BC, Greene P, Boyer HW, Grable J, Rosenberg JM. Structure of the DNA-Eco RI endonuclease recognition complex at 3 A resolution. Science. 1986 Dec 19;234(4783):1526-41.

Mulligan EA, Hatchwell E, McCorkle SR, Dunn JJ. Differential binding of Escherichia coli McrA protein to DNA sequences that contain the dinucleotide m5CpG. Nucleic Acids Res. 2010 Apr;38(6):1997-2005.

Neaves KJ, Cooper LP, White JH, Carnally SM, Dryden DT, Edwardson JM, Henderson RM. Atomic force microscopy of the EcoKI Type I DNA restriction enzyme bound to DNA shows enzyme dimerization and DNA looping. Nucleic Acids Res. 2009 Apr;37(6):2053-63.

O'Driscoll J, Glynn F, Cahalane O, O'Connell-Motherway M, Fitzgerald GF, Van Sinderen D. Lactococcal plasmid pNP40 encodes a novel, temperature-sensitive restriction-modification system. Appl Environ Microbiol. 2004 Sep;70(9):5546-56.

O'Driscoll J, Fitzgerald GF, van Sinderen D. A dichotomous epigenetic mechanism governs expression of the LlaJI restriction/modification system. Mol Microbiol. 2005 Sep;57(6):1532-44.

O'Driscoll J, Heiter DF, Wilson GG, Fitzgerald GF, Roberts R, van Sinderen D. A genetic dissection of the LlaJI restriction cassette reveals insights on a novel bacteriophage resistance system. BMC Microbiol. 2006 Apr 28;6:40.

Oppenheim AB, Kobiler O, Stavans J, Court DL, Adhya S. Switches in bacteriophage lambda development. Annu Rev Genet. 2005;39:409-29.

Orlowski J, Bujnicki JM. Structural and evolutionary classification of Type II restriction enzymes based on theoretical and experimental analyses. Nucleic Acids Res. 2008 Jun;36(11):3552-69.

O'Sullivan DJ, Zagula K, Klaenhammer TR. In vivo restriction by LlaI is encoded by three genes, arranged in an operon with llaIM, on the conjugative Lactococcus plasmid pTR2030. J Bacteriol. 1995 Jan;177(1):134-43.

Panne D, Raleigh EA, Bickle TA. The McrBC endonuclease translocates DNA in a reaction dependent on GTP hydrolysis. J Mol Biol. 1999 Jul 2;290(1):49-60.

Panne D, Müller SA, Wirtz S, Engel A, Bickle TA. The McrBC restriction endonuclease assembles into a ring structure in the presence of G nucleotides. EMBO J. 2001 Jun 15;20(12):3210-7.

Parma DH, Snyder M, Sobolevski S, Nawroz M, Brody E, Gold L. The Rex system of bacteriophage lambda: tolerance and altruistic cell death. Genes Dev. 1992 Mar;6(3):497-510.

Pieper U, Brinkmann T, Krüger T, Noyer-Weidner M, Pingoud A. Characterization of the interaction between the restriction endonuclease McrBC from E. coli and its cofactor GTP. J Mol Biol. 1997 Sep 19;272(2):190-9.

Pieper U, Schweitzer T, Groll DH, Gast FU, Pingoud A. The GTP-binding domain of McrB: more than just a variation on a common theme? J Mol Biol. 1999 Sep 24;292(3):547-56.

Pieper U, Pingoud A. A mutational analysis of the PD...D/EXK motif suggests that McrC harbors the catalytic center for DNA cleavage by the GTP-dependent restriction enzyme McrBC from Escherichia coli. Biochemistry. 2002 Apr 23;41(16):5236-44.

Pieper U, Groll DH, Wünsch S, Gast FU, Speck C, Mücke N, Pingoud A. The GTP-dependent restriction enzyme McrBC from Escherichia coli forms high-molecular mass complexes with DNA and produces a cleavage pattern with a characteristic 10-base pair repeat. Biochemistry. 2002 Apr 23;41(16):5245-54.

Raghavendra NK, Bheemanaik S, Rao DN. Mechanistic insights into type III restriction enzymes. Front Biosci (Landmark Ed). 2012 Jan 1;17:1094-107.

Raleigh EA, Wilson G. Escherichia coli K-12 restricts DNA containing 5-methylcytosine. Proc Natl Acad Sci U S A. 1986 Dec; 83(23): 9070–9074.

Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, Blumenthal RM, Degtyarev SKh, Dryden DT, Dybvig K, Firman K, Gromova ES, Gumport RI, Halford SE, Hattman S, Heitman J, Hornby DP, Janulaitis A, Jeltsch A, Josephsen J, Kiss A, Klaenhammer TR, Kobayashi I, Kong H, Krüger DH, Lacks S,Marinus MG, Miyahara M, Morgan RD, Murray NE, Nagaraja V, Piekarowicz A, Pingoud A, Raleigh E, Rao DN, Reich N, Repin VE, Selker EU, Shaw PC, Stein DC, Stoddard

BL, Szybalski W, Trautner TA, Van Etten JL, Vitor JM, Wilson GG, Xu SY. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. Nucleic Acids Res. 2003 Apr 1;31(7):1805-12.

Rosamond J, Endlich B, Linn S. Electron microscopic studies of the mechanism of action of the restriction endonuclease of Escherichia coli B. J Mol Biol. 1979 Apr 25;129(4):619-35.

Scheffzek K, Ahmadian MR, Kabsch W, Wiesmüller L, Lautwein A, Schmitz F, Wittinghofer A. The Ras-RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants. Science. 1997 Jul 18;277(5324):333-8.

Siwek W, Czapinska H, Bochtler M, Bujnicki JM, Skowronek K. Crystal structure and mechanism of action of the N6-methyladenine-dependent type IIM restriction endonuclease R.*Dpn*I. Nucleic Acids Res. 2012 Aug;40(15):7563-72.

Snyder L, McWilliams K. The rex genes of bacteriophage lambda can inhibit cell function without phage superinfection. Gene. 1989 Sep 1;81(1):17-24.

Snyder L. Phage-exclusion enzymes: a bonanza of biochemical and cell biology reagents? Mol Microbiol. 1995 Feb;15(3):415-20. Review.

Stewart FJ, Raleigh EA. Dependence of McrBC cleavage on distance between recognition elements. Biol Chem. 1998 Apr-May;379(4-5):611-6.

Stewart FJ, Panne D, Bickle TA, Raleigh EA. Methyl-specific DNA binding by McrBC, a modification-dependent restriction enzyme. J Mol Biol. 2000 May 12;298(4):611-22.

Sukackaite R, Grazulis S, Tamulaitis G, Siksnys V. The recognition domain of the methyl-specific endonuclease McrBC flips out 5-methylcytosine. Nucleic Acids Res. 2012 Aug;40(15):7552-62.

Summers WC. The strange history of phage therapy. Bacteriophage. 2012 Apr 1;2(2):130-133.

Sutherland E, Coe L, Raleigh EA. McrBC: a multisubunit GTP-dependent restriction endonuclease. J Mol Biol. 1992 May 20;225(2):327-48.
Tesfazgi Mebrhatu M, Wywial E, Ghosh A, Michiels CW, Lindner AB, Taddei F, Bujnicki JM, Van Melderen L, Aertsen A. Evidence for an evolutionary antagonism between Mrr and Type III modification systems. Nucleic Acids Res. 2011 Aug;39(14):5991-6001.

Tesmer JJ, Berman DM, Gilman AG, Sprang SR. Structure of RGS4 bound to AlF4--activated G(i alpha1): stabilization of the transition state for GTP hydrolysis. Cell. 1997 Apr 18;89(2):251-61.

Vallee BL, Auld DS. Zinc coordination, function, and structure of zinc enzymes and other proteins. Biochemistry. 1990 Jun 19;29(24):5647-59.

van Noort J, van der Heijden T, Dutta CF, Firman K, Dekker C. Initiation of translocation by Type I restriction-modification enzymes is associated with a short DNA extrusion. Nucleic Acids Res. 2004 Dec 14;32(22):6540-7. Print 2004.

Waite-Rees PA, Keating CJ, Moran LS, Slatko BE, Hornstra LJ, Benner JS. Characterization and expression of the Escherichia coli Mrr restriction system. J Bacteriol. 1991 Aug;173(16):5207-19.

Wang H, Guan S, Quimby A, Cohen-Karni D, Pradhan S, Wilson G, Roberts RJ, Zhu Z, Zheng Y. Comparative characterization of the PvuRts1I family of restriction

enzymes and their application in mapping genomic 5-hydroxymethylcytosine. Nucleic Acids Res. 2011 Nov;39(21):9294-305.

Winkler FK, Banner DW, Oefner C, Tsernoglou D, Brown RS, Heathman SP, Bryan RK, Martin PD, Petratos K, Wilson KS. The crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. EMBO J. 1993 May;12(5):1781-95.

Wittebole X, De Roock S, Opal SM. A historical overview of bacteriophage therapy as an alternative to antibiotics for the treatment of bacterial pathogens. Virulence. 2014 Jan 1;5(1):226-35.

Xiao CL, Zhu S, He M, Chen, Zhang Q, Chen Y, Yu G, Liu J, Xie SQ, Luo F, LiangZ, Wang DP, Bo XC, Gu XF, Wang K, Yan GR. N(6)-Methyladenine DNA Modification in the Human Genome. Mol Cell. 2018 Jul 19;71(2):306-318.e7.

Yang X, Xu M, Yang ST. Restriction modification system analysis and development of in vivo methylation for the transformation of Clostridium cellulovorans. Appl Microbiol Biotechnol. 2016 Mar;100(5):2289-99.

Yu YT, Snyder L. Translation elongation factor Tu cleaved by a phage-exclusion system. Proc Natl Acad Sci U S A. 1994 Jan 18;91(2):802-6.

Yuan R, Hamilton DL, Burckhardt J. DNA translocation by the restriction enzyme from E. coli K. Cell. 1980 May;20(1):237-44.

Zheng Y, Cohen-Karni D, Xu D, Chin HG, Wilson G, Pradhan S, Roberts RJ. A unique family of Mrr-like modification-dependent restriction endonucleases. Nucleic Acids Res. 2010 Sep;38(16):5527-34.

Chapter 2. The crystal structure of the *Helicobacter pylori* LlaJI.R1 N-terminal domain provides a model for site-specific DNA binding

The crystal structure of the *Helicobacter pylori* LlaJI.R1 N-terminal domain provides a model for site-specific DNA binding

Christopher J. Hosford[1] and Joshua S. Chappie[1,*]

From the [1]Department of Molecular Medicine, Cornell University, Ithaca NY 14853

Running title: *Structure of Helicobacter pylori LlaJI.R1 N-terminal domain*

*To whom correspondence should be addressed: Joshua S. Chappie: Department of Molecular Medicine, Cornell University, Ithaca NY 14853; chappie@cornell.edu; Tel. (607) 253-3654; Fax. (607) 253-3659.

Key words: LlaJI, restriction system, B3 domain, DNA binding, McrB, asymmetric dimerization

ABSTRACT

Restriction modification systems consist of an endonuclease that cleaves foreign DNA site-specifically and an associated methyltransferase that protects the corresponding target site in the host genome. Modification-dependent restriction systems, in contrast, specifically recognize and cleave methylated and/or glucosylated DNA. The LlaJI restriction system contains two 5-methyl-cytosine (5mC) methyltransferases (LlaJI.M1 and LlaJI.M2) and two restriction proteins (LlaJI.R1 and LlaJI.R2). LlaJI.R1 and LlaJI.R2 are homologs of McrB and McrC respectively, which in *Escherichia coli* function together as a modification-dependent restriction complex specific for 5mC-containing DNA. *Lactococcus lactis* LlaJI.R1 binds DNA site-specifically, suggesting that the LlaJI system uses a different mode of substrate recognition. Here we present the structure of the N-terminal DNA binding domain of *Helicobacter pylori* LlaJI.R1 at 1.97Å resolution, which adopts a B3 domain fold. Structural comparison to B3 domains in plant transcription factors and other restriction enzymes identifies key recognition

motifs responsible for site-specific DNA binding. Moreover, biochemistry and structural modeling provide a rationale for how *Helicobacter pylori* LlaJI.R1 may bind a target site that differs from the five base pair sequence recognized by other LlaJI homologs and identify residues critical for this recognition activity. These findings underscore the inherent structural plasticity of B3 domains, allowing recognition of a variety of substrates using the same structural core.

INTRODUCTION

Classical restriction modification (RM) systems are ubiquitous in bacteria and act as a requisite layer of defense against predatory bacteriophage viruses (1). These systems consist of a restriction endonuclease and a methyltransferase, which provide the dual-function of cleaving exogenous DNA site-specifically while protecting the host genome via methylation of the corresponding recognition sequence (2). Three variants of RM systems – type I, type II, and type III – have been identified and differ in their structural composition and mechanism of restriction. Type I systems are multifunctional complexes containing separate restriction, methylation, and DNA-sequence recognition subunits. These machines require Mg^{2+} and ATP, catalyze both restriction and methylation, and cut DNA non-specifically far from their recognition sites (3). Type II systems are the simplest, generally existing as dimers that carry out the recognition and restriction activities. These enzymes do not require ATP and have a separate, associated methyltransferase (4). Homodimeric type II restriction enzymes recognize DNA sequences that are symmetric while those that are heterodimeric can bind asymmetric sequences (5). Type III systems contain separate modification (Mod) and restriction

(Res) subunits that form homodimeric Mod2 and a heterotetrameric Res2Mod2 complexes and catalyze both restriction and methylation in a Mg^{2+} and ATP-dependent manner (6). They differ from type I systems, however, in that they require two inversely oriented recognitions sites that can vary in their spatial separation (7).

Modification-dependent restriction systems (MDRS) – colloquially referred to as type IV systems – recognize and cleave modified DNA (8). McrA and McrBC are prototypical MDRSs that target DNA containing 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) (9, 10, 11, 12, 13, 14). McrA is a small, dimeric protein that recognizes the symmetrically methylated sequence Y5mCRG (15). The McrB and McrC proteins together form a conserved, two-component restriction complex capable of long range DNA translocation similar to type I and type III enzymes. *Escherichia coli* (Ec) McrB contains an N-terminal DNA binding domain (12) and a C-terminal AAA+ motor domain that hydrolyzes GTP and mediates nucleotide-dependent oligomerization into heptameric rings (16). McrB's basal GTPase activity is stimulated via interaction with its partner endonuclease McrC (13), which cannot bind DNA on its own and *in vitro* only associates with the McrB oligomer (17). Biochemical studies suggest a model for DNA cleavage in which McrB and McrC assemble at two distant R^MC sites (where R is a purine, and ^MC is a methylcytosine) and translocate in a manner that requires stimulated GTP hydrolysis (10, 18). Collision of McrBC complexes triggers cleavage of both DNA strands close to one of the R^MC sites (14, 19). Other MDRS families display a variable spectrum of specificity for different modifications. These include MspJI, which recognizes 5mC and 5hmC (20), the PvuRts1I family, whose members show unique individual specificities for 5hmC and/or 5-

glucosylhydroxymethylcytosine (5ghmC) (21), and GmrSD, which recognizes 5ghmC (22). Structural studies of McrB, MspJI, PvuRtsI, and *Aba*SI suggest type IV systems employ a generalized base-flipping mechanism for recognition of the modified DNA (23, 24, 25, 26, 27).

The LlaJI restriction cassette was first identified in *Lactococcus lactis* on the naturally occurring plasmid pNP40 and shown to confer resistance against common lactococcal phages (28). It consists of an operon encoding two 5mC-methyltransferases, LlaJI.M1 and LlaJI.M2, and two restriction proteins, LlaJI.R1 and LlaJI.R2, both of which are absolutely required for restriction activity *in vivo* (29). The M1 and M2 methyltransferase activities modulate expression of LlaJI operon *in vivo* (30). Although formally classified as a type II R/M system (REBASE enzyme # 10100, NEB), LlaJI.R1 and LlaJI.R2 share domain homology with McrB and McrC respectively. R1 contains sequence motifs that identify its C-terminal portion as a GTP-specific AAA+ domain and R2 contains a conserved C-terminal PD-(D/E)xK endonuclease domain. These features suggest LlaJI enzymes function more like McrBC than other Type II systems.

Unlike McrB, however, *L. lactis* LlaJI.R1 binds DNA site-specifically, recognizing the asymmetric 5'-GACGC-3' sequence in one strand and 5'-GCGTC-3' in the other strand (29). Other LlaJI homologs have been identified in *Helicobacter pylori*, *Streptococcus pyogens*, *Bacillus cereus*, and *Clostridium cellulovorans* (29, 31). Of these, *Clostridium* cellulovorans LlaJI has also been shown to target the same asymmetric, five base pair sequence (Yang, et al. 2016). How LlaJI proteins recognize DNA site-specifically is unknown. Here we present the structure of the N-terminal DNA binding domain of *Helicobacter pylori* LlaJI.R1 (HpR1Δ136) at 1.97 Å resolution,

which adopts a B3 domain fold. Structural comparison to B3 domain-containing plant transcription factors and restriction endonucleases identifies the key recognition motifs responsible for site-specific DNA binding. Additional evidence from biochemistry and structural modeling argues that HpLlaJI.R1 binds a target site that differs from the five base pair sequence recognized by the *Lactococcus lactis* and *Clostridium cellulovorans* LlaJI homologs. Mutagenesis further identifies residues R17 and R60 as critical determinants of HpLlaJI.R1 DNA binding. Together, these findings underscore the inherent structural plasticity previously noted for B3 domains, which confers specificity to different sequences via the same structural core.

RESULTS

Structure and topology of the HpLlaJI.R1 N-terminal domain

Though previous studies show LlaJI.R1 binds DNA site-specifically (29, 31), the molecular means through which this is achieved remains unknown. Numerous attempts to purify either the full-length *Lactococcus lactis* LlaJI.R1 or its isolated N-terminal DNA binding domain for structural studies were unsuccessful. Bioinformatics identifies various other species harboring the LlaJI operon, including *Helicobacter pylori* (Hp) (29). Computational analyses of these homologs by fold matching and structural prediction algorithms failed to identify a reliable template for modeling DNA interactions. To understand how the LlaJI.R1 binds DNA site-specifically, we therefore crystallized the N-terminal domain of HpLlaJI.R1 (HpR1Δ136) and determined its structure at 1.97 Å by selenium SAD phasing (32) (Fig. 1).

HpR1Δ136 crystallizes in the space group P1 with four molecules (A-D) in the

asymmetric unit organized as two homodimers packed end to end, with molecules A and B and molecules C and D pairing together (Fig. 1A). These dimers superimpose with an RMSD of 0.589Å. Each HpR1Δ136 monomer consists of a core six-stranded β sheet that folds into a pseudo-beta barrel flanked on four separate edges by alpha helices (α1-α4) (Fig. 1B,C). An additional β-strand (β7) inserts at the dimer interface and breaks the symmetry, adopting an antiparallel configuration with $\beta1^{B}/\beta1^{D}$ and a parallel configuration with $\beta1^{A}/\beta1^{C}$ (Fig. 1B,C). Clear connectivity between β7 and α4 can be traced in molecule B (Fig. 1D). Structural superposition of the two asymmetric dimers suggests β7 is connected in the same manner in molecule D despite the lack of density for the α4-β7 loop (Fig. 1E). We observe no density for the corresponding β7 strands in either molecule A or molecule C.

β7 residues L127 and F129 interact with a hydrophobic cluster sandwiched between β1 and the amphipathic α2 helix in each monomer (Fig. 2A). I24, H27, and F28 in α2 and V115, L116, and L119 in α4 provide additional stabilizing contacts across the dimer interface (Fig. 2A,B). β7 insertion helps space these elements and prevent steric clashing that otherwise would occur. A total interaction surface of 800 Å2 is shared between the monomers. Size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS) indicates HpR1Δ136 dimerizes in solution (Fig. 2C), suggesting the observed molecular organization in the crystal lattice is not simply a packing artifact. Deletion of β7 renders HpR1Δ136 insoluble. Point mutations at the dimer interface in α2 (I24N, H27E, F28E) and α4 (L116E and L199E) displayed similar insolubility phenotypes and could not be purified. Only the V115N mutant retained

solubility and, like wildtype, forms a stable dimers in solution when analyzed by SEC-MALS (Fig. 3A). These data support the notion that HpR1Δ136 dimerization is required for stability and biologically relevant.

HpR1Δ136 adopts a B3 domain fold

The DALI alignment algorithm (33) identifies several structural homologs of HpR1Δ136. These include the DNA binding domain of the *Arabidopsis thaliana* (At) auxin-dependent transcription factor ARF1 (PDB ID: 4ldx, Z-score = 8.6, RMSD = 3.0), the C-terminal fragment of the BfiI restriction endonuclease (PDB ID: 3zi5, z-score = 9.7, RMSD = 2.1), and the N-terminal fragment of the EcoRII restriction endonuclease (PDB ID: 3hqf, z-score = 8.1, RMSD = 2.7). Each is comprised of a B3 domain (34, 35, 36, 37, 38). B3 domains share a common pseudo-barrel architecture (SCOP number 101935) and act as recognition modules that bind DNA site-specifically (39). Structural superposition confirms HpR1Δ136 similarly adopts a B3 domain fold (Fig. 4A, Fig. S1). Importantly, this fold is structurally distinct from the analogous region in EcMcrB, which preferentially binds DNA containing methylated cytosines.

B3 domains contain two critical regions that confer DNA target site specificity. These recognition motifs, termed the N-arm and C- arm, reside on opposite edges of the pseudo-barrel core and form a wrench-like structure that contacts the major groove (36, 37, 40). In this arrangement, the N-arm specifically associates with the 5'-half of the target site and the C-arm engages the 3'-half. Comparison to the DNA-bound AtARF1 structure identifies these key features within the HpR1Δ136 model: the N-arm encompasses the β1-β2 loop and the α1 helix and the C-arm localizes to the β3-β4 loop

(Fig. 1C, Fig. 4B). Both proteins display a comparable electrostatic surface, with an extensive basic patch positioned between the N- and C-arms and coincident with the DNA binding face of AtARF1 (Fig. 4C). This mode of substrate binding is conserved among DNA-bound B3 domain structures and is consistent with other HpR1Δ136 structural superpositions (Fig. S1). An exception is the B3 domain of NgoAVII, whose orientation on DNA is inverted such that the N-arm associates with the 3' half of the target site and the C-arm with the 5' half (40).

The BfiI contains other unique motifs (termed the N- and C-loops) that provide additional phosphate backbone and minor groove interactions (37). These are shortened in EcoRII and absent in all previously characterized plant B3 domains (36, 38, 41, 42, 43). HpR1Δ136 similarly lacks these segments, suggesting it either evolved from a more simplified common ancestor or lost these segments over time due to a lack of selective pressure.

The putative DNA binding surface of each HpR1Δ136 monomer faces away from the dimer interface, suggesting that HpR1Δ136 has the capacity to bind two DNA target sites simultaneously. The asymmetric orientation of the β7 strand, however, positions E131 close to one of the binding sites and alters its surface charge potential in a manner that makes it less basic (Fig. 4D). This intrinsic difference would allow one monomer to bind more efficiently and could bias the arrangement of HpLlaJI on DNA.

All previously characterized B3 domains exist as monomers (36, 37, 38, 39, 40, 41, 42, 43, 44). To understand what hinders dimerization in these contexts, we superimposed the coordinates of other B3 domains onto our HpR1Δ136 dimer and examined the orientation of secondary structure features relative to the dimer interface

(Fig. 5). Although EcoRII and BfiI have structurally equivalent β-strands that partially align with β7, they also contain helical segments that sterically prevent two monomers from coming together (Fig. 5A,B). The α1 helices of VRN1 and UbaLAI would similarly clash and block dimerization (Fig. 5C, D). AtARF1, NgoAVII, RAV1, and At1g16640.1, in contrast, all lack a corresponding β7 strand (Fig. 5E-H), suggesting monomers cannot be properly spaced to avoid collision. The stabilizing hydrophobic interactions provided by β7 would also be absent. These observations highlight the importance of secondary structure features in modulating the oligomeric state of B3 domains and will be useful for predicting interactions in other uncharacterized proteins that contain this conserved fold.

HpR1Δ136 structure provides model for site specific binding

Previous biochemical and genetic studies indicate that *Lactococcus lactis* and *Clostridium cellulovorans* LlaJI target the five base pair sequence 5'-GACGC-3' (29, 31). HpR1Δ136 shows weak affinity for DNA containing this sequence (Ll) when assessed by filter binding (Fig. 6A, blue). Scrambling the putative binding sequence (Llscr) has no effect on the affinity (Fig. 6A, green), suggesting this represents the basal level for non-specific DNA binding by HpR1Δ136. EcMcrB, in contrast, preferentially binds 5mC-containing DNA (5mC) but does not bind a non-methylated version of the same substrate (nm) under the same assay conditions (Fig. 6A, black versus red). EcMcrB similarly does not bind either the Ll or Llscr substrates (Fig. 6A, yellow and orange), underscoring how its binding depends on the presence of methylated cyotsines.

Unexpectedly, HpR1Δ136 showed enhanced affinity for the *E. coli* specific 5mC

and nm substrates (Fig. 6A, light blue and purple) relative Ll and Llscr substrates. We attribute this to subtle sequence differences as the binding is independent of methylation status. The 5mC and nm substrates likely contain sequence fragments that more closely mimic that preferred recognition site of HpR1.LlaJI, which is distinct from both EcMcrB and other LlaJI homologs.

Despite a common fold, B3 domains exhibit divergent sequence preferences. Previous structural and biochemical data show that the C-arm length can influence the length of the recognized target site (Fig. 6B). A longer C-arm is excluded from the major groove (Fig. 6C), decreasing the overall binding footprint and biasing recognition towards a five base pair site (36, 37). A shorter C-arm affords greater access to the DNA bases, which in some instances increases the number of specific contacts and extends the target site to six bases (37, 38). The amino acid composition of the N- and C-arms ultimately dictates specificity, however, and thus some B3 domains with shorter C-arms still bind five base pair sites (40, 44). Structural superposition reveals shorter C-arm in HpR1Δ136 (Fig. 6C).

In the absence of a DNA bound complex and without explicit knowledge of the HpLlaJI target site, we used the AtARF1-DNA structure as a proxy to identify side chains that might contribute to specificity. The N-arm residue H136 and C-arm residues R181, P184, and R186 are critical for AtARF1 DNA binding (Fig. 6D). Structural modeling reveals similar residues in HpR1Δ136 (Fig. 6E), with H14 and R17 in the N-arm and P59 and R60 in the C-arm poised to provide base-specific contacts. Interestingly, R17 is spatially oriented like R181 in AtARF1, hinting that it would contact the 3'-half of the target site despite being localized in the N-arm.

To confirm the significance of our structural modeling, we mutated the predicted binding residues in HpR1Δ136 and assessed how each substitution affects interaction with the *E. coli* specific nm DNA substrate via filter binding (Fig 6F). H14A (red), R17A (orange), and R60A (light blue) mutations show a marked decrease in affinity for nm DNA versus wildtype (WT, purple), while P59A (green) shows less of an effect. The R17A/R60A double mutant (brown) completely abolishes binding. This finding was corroborated using electrophoretic mobility shift assays (EMSAs) to measure the association of HpR1Δ136 with digested, non-methylated λ-phage DNA (Fig. 7). We observe a significant gel shift with wildtype HpR1Δ136. H14A and P59A show similar shifts in this assay while the individual R17A and R60A substitutions produce a moderate reduction in binding. The R17A/R60A double mutant, however, significantly impairs binding (Fig. 7), similar to its effects on the nm DNA substrate in the filter binding assay (Fig. 6F). All of these mutants form stable dimers in solution (Fig. 3B), arguing that their effects are not due to global structural perturbations. Together these data implicate R17 and R60 as critical determinants of HpR1Δ136 DNA binding.

B3 domains contain conserved residues that associate with 'clamp' phosphates at the 5' ends of each strand in the target site (37). R177 and K126 form these interactions in AtARF1 (R81 and K23 in EcoRII; R272 and K340 in BfiI; K27 and K82 in UbaLAI; K212 and K275 in R.NgoAVII). In HpR1Δ136, R6 is poised to act on one strand while K50 could perform a similar function on the opposing strand. K50 is positioned away from the modelled DNA backbone in the apo state and may be reoriented upon target recognition. Conformational rearrangements in the BfiI and R.NgoAVII B3 domains have previously been observed upon DNA binding (35, 37,

40).

DISCUSSION

Here we described the structure of the HpLlaJI.R1 DNA binding domain and demonstrated that it adopts a B3 domain fold. B3 domains are prevalent among bacterial restriction endonucleases and plant transcription factors, where they function as site-specific DNA binding modules (37, 39, 45). Previous crystallographic studies revealed that the N- and C-arms determine the specificity of each individual B3 domain and confer structural plasticity to the conserved core scaffold (36, 37, 38, 40, 43). Our structural data and modeling identifies the N- and C-arms in HpR1Δ136 along with key residues that likely form direct contacts with the DNA backbone, clamp phosphates, and specific bases. HpR1Δ136 has weak affinity for DNA containing the asymmetric five base pair site that other LlaJI homologs target (29, 31) and a surprisingly stronger affinity for the EcMcrB-specific DNA substrates, regardless of their methylation status (Fig. 6A). We note that HpR1Δ136 contains a shorter C-arm and thus could potentially bind a six base pair site. While further studies are required to pinpoint the target site of HpLlaJI.R1, our findings offer a general model for site-specific binding and provide a structural explanation for why LlaJI homologs do not target modifications despite sharing a similar domain organization with McrBC.

Importantly, we identify the N-arm R17 and C-arm R60 residues as critical determinants of DNA binding and specificity in HpR1Δ136. Individual point mutations at these positions display moderate defects while a combined double mutant completely abolishes DNA binding in all assays tested (Figs. 6F and 7). The H14A and P59A

mutations show varying effects depending on the specific substrate used and the sensitivity of the assay. While H14 and P59 may also impart specific binding interactions, their contribution is likely context dependent.

HpLlaJI.R1 is unique in that its isolated B3 domain dimerizes, both in solution (Fig. 2C) and *in crystallo* (Fig. 1). β7 is absolutely essential for HpR1Δ136 dimerization and structural stability, as a truncation of this motif renders the protein insoluble. Our structure shows that direct dimerization of other B3 domains is hindered by either (i) the intrinsic lack of a structurally equivalent β7 strand or (ii) the presence of additional helical motifs at the N- or C-terminus that sterically clash with β7 or α4 at the dimer interface (Fig. 5). Dimerization of other B3 domain-containing proteins instead occurs through additional structural elements. For instance, AtARF1 monomers associate through a separate dimerization domain, which facilitates cooperative binding of the B3 domains to two anti-parallel 5'-TGTCTC-3' sites on opposing strands (38). BfiI, EcoRII, and R.NgoAVII also dimerize but through their respective nuclease domains (34, 35, 40). These observations will help in classifying uncharacterized B3 domains and predicting their architectural organization.

Our structural data show that the β7 strand from one HpR1Δ136 monomer is asymmetrically stabilized at the dimer interface while the corresponding region in the other monomer remains disordered. The orientation of this strand dictates the electrostatic landscape on the dimer surface, making the DNA binding site in one monomer more basic than the other. While we cannot completely rule out that this is an artifact of crystallization, an analogous phenomenon was noted in the rotavirus A non-structural protein 3 (NSP3) homodimer (46). There the asymmetric stabilization of a

helix from one monomer creates a single positively charged binding site that ultimately leads to a stoichiometry of 2:1 NSP3:viral mRNA (46). We speculate that HpLlaJI.R1 may bind DNA with a similar 2:1 stoichiometry, but that only one B3 domain will directly contact the target site.

Asymmetric binding could have important implications for the assembly of a cleavage-competent LlaJI restriction complex. Like McrB, LlaJI.R1 contains a conserved GTP-specific AAA+ domain at its C-terminus (29). EcMcrB forms heptameric rings in the presence of GTP and this oligomerization is critical for recruiting its partner endonuclease McrC (16), which cannot bind DNA on its own and preferentially associates with the assembled AAA+ domain (17). While the exact organization of McrBC on DNA has yet to be elucidated, biochemical and structural studies have shown the EcMcrB N-terminal domain binds a single methylated cytosine via base flipping (23). The intrinsic asymmetry of the McrBC complex therefore imposes constraints on how the individual subunits interact with the R^MC site and suggests that some monomers are directly engaged while others are not. The asymmetric HpR1Δ136 dimer could reflect a similar structural constraint in the LlaJI system wherein the alternative positioning of the β7 strand dictates which monomers bind the target sequence. Further structural characterization of both systems will be necessary to parse out how substrate binding, GTP-dependent assembly, and nuclease recruitment are coordinated in each case. Although LlaJI and McrBC differ in their specificity and targeting, we predict the general molecular mechanisms governing the function of LlaJI and McrBC will be conserved.

EXPERIMENTAL PROCEDURES

Cloning, expression and purification HpLlaJI.R1 constructs

DNA encoding the *Helicobacter pylori* LlaJI.R1 protein (DOE IMG/M ID 637022177) was codon optimized for *E. coli* expression and synthesized commercially by Bio Basic Inc. DNA encoding the N-terminal domain (HpR1Δ136; residues 1-136) was amplified by PCR and cloned into pET21b, introducing a 6xHis tag at the C-terminus. Selenomethionine labeled (SeMet) HpR1Δ136 was expressed in minimal media using methionine auxotrophs (T7 Express Crystal Competent E. coli, New England Biolabs) according to manufacturer protocols. Native HpR1Δ136 was transformed into BL21(DE3) cells, grown at 37°C in Terrific Broth to an OD600 of 1.0, and then induced with 0.3 mM IPTG overnight at 19°C. All cells were harvested, washed with nickel load buffer (20 mM HEPES pH 7.5, 500 mM NaCl, 30 mM imidazole, 5% glycerol (v:v), and 5 mM β-mercaptoethanol), and pelleted a second time. Pellets were typically flash frozen in liquid nitrogen and stored at -80°C.

Thawed pellets from 500 ml cultures were resuspended in 30 ml of nickel load buffer supplemented with 10 mM PMSF, 5 mg DNAse (Roche), 5 mM MgCl2, and a Roche complete protease inhibitor cocktail tablet (Roche). Lysozyme was added to 1 mg/ml and the mixture was incubated for 15 minutes rocking at 4°C. Cells were disrupted by sonication and the lysate was cleared of debris by centrifugation at 13 000 rpm (19 685 g) for 30 minutes at 4°C. For native and SeMet HpR1Δ136, the supernatant was filtered, loaded onto a 5 ml HiTrap chelating column charged with NiS04 and then washed with nickel load buffer. HpΔ136 was eluted with an imidazole gradient from 30 mM to 1 M. Pooled fractions were dialyzed overnight at 4°C into SP loading buffer (20

mM HEPES pH 7.5, 50 mM NaCl, 1 mM EDTA, 5% glycerol (v:v), and 5 mM DTT). The sample was applied to a 5 ml HiTrap SP HP column equilibrated with SP loading buffer and then washed with SP loading buffer. HpR1Δ136 was eluted with a NaCl gradient from 50 mM to 1 M. Pooled fractions were concentrated and further purified by size exclusion chromatography (SEC) using a Superdex 200 10/300 column. All proteins were exchanged into a final buffer of 20mM HEPES pH 7.5, 150mM KCl, 5 mM MgCl2, and 1mM DTT (5mM for SeMet labelled) during SEC and concentrated to 5-40 mg/ml. Concentrations of purified proteins were determined by SDS-PAGE and densitometry compared against BSA standards. All amino acid substitutions were introduced into HpR1Δ136 in pET21b by Quikchange PCR and mutant proteins were purified as described for wildtype.

Cloning, expression and purification EcMcrB

DNA encoding the Escherichia coli McrB protein (Uniprot P15005) was codon optimized for *E. coli* expression and synthesized commercially by GENEART. DNA encoding the full-length protein (EcMcrB FL) was amplified by PCR and cloned into c2xP, a modified pMAL c2x vector with an HRV3C protease site replacing the Factor Xa site directly upstream of the *mcrB* gene. Native EcMcrB FL expressed as N-terminal maltose binding protein fusion in BL21(DE3) cells. Transformed cells were grown at 37°C in Terrific Broth to an OD600 of 0.8-1.0, and then induced with 0.3 mM IPTG overnight at 19°C. All cells were harvested, washed with amylose loading buffer (20 mM Tris-HCl pH 8.0, 500 mM NaCl, 5% glycerol (v:v), 1 mM EDTA, and 1 mM DTT), and pelleted a second time. Pellets were typically flash frozen in liquid nitrogen and

stored at -80°C.

Thawed pellets from 500 ml cultures were resuspended in 30 ml amylose loading buffer supplemented with 10 mM PMSF, 5 mg DNAse (Roche), 5 mM MgCl2, and a Roche complete protease inhibitor cocktail tablet (Roche). Lysozyme was added to 1 mg/ml and the mixture was incubated for 15 minutes rocking at 4°C. Cells were disrupted by sonication and the lysate was cleared of debris by centrifugation at 13 000 rpm (19 685 g) for 30 minutes at 4°C. The supernatant was filtered, loaded onto 40 ml of packed amylose resin (New England Biolabs) and then washed with amylose loading buffer. EcMcrB FL was eluted with amylose loading buffer supplemented with 10 mM D-maltose. Hrv3C protease was added to pooled fractions and dialyzed overnight at 4°C into Q loading buffer (20 mM Tris-HCl pH 8.0, 50 mM NaCl, 1 mM EDTA, 5% glycerol (v:v), and 1 mM DTT). The sample was applied to a 5 ml HiTrap Q HP column equilibrated with Q loading buffer and then washed with Q loading buffer. EcMcrB FL was eluted with a NaCl gradient from 50 mM to 1 M. Pooled fractions were concentrated and further purified by size exclusion chromatography (SEC) using a Superdex 75 10/30 pg column. All proteins were exchanged into a final buffer of 20mM HEPES pH 7.5, 150mM KCl, 5 mM MgCl2, and 1mM DTT during SEC and concentrated to 5-40 mg/ml.

Crystallization, X-ray data collection, and structure determination

SeMet HpR1Δ136 was crystallized by sitting drop vapor diffusion in 0.1 M MMT pH6.5, 25% PEG 1,500 (v:v) by mixing 1 μL of protein with 1 μL of the condition with a final drop size of 2 μL and reservoir volume of 65 μL. Crystals appeared within 2-8

days at 20°C. Samples were cryoprotected with Parabar 10312 and frozen in liquid nitrogen. Single-wavelength anomalous diffraction (SAD) data of two crystals were collected remotely on the tuneable NE-CAT 24-ID-C beamline at the Advanced Photon Source at the selenium edge energy at 12.663 kEv (Table 1). Crystal 1 was of the space group P1 with unit cell dimensions a = 37.47, b = 44.39, c = 84.78 Å and $\alpha = 98.04°$, $\beta = 94.37°$, $\gamma = 98.52°$ and showed strong anomalous signal. Crystal 2 was of the space group P1 with unit cell dimensions a = 37.53, b = 43.77, c = 85.09 Å and $\alpha = 97.87°$, $\beta = 93.86°$, $\gamma = 97.77°$. Both crystals were prepared in the same condition but exhibited mosaicities of 0.20871° and 0.11033° respectively. Data were integrated and scaled using XDS (47) and AIMLESS (48) via the NE-CAT RAPD pipeline. Se-SAD phasing with the data from crystal 1 yielded an initial model that was incomplete and contained a few regions of ambiguity. Heavy atom sites were located using SHELX (49) and phasing, density modification, and initial model building was carried out using the Autobuild routines of the PHENIX package (50). Further cycles of model building and refinement was carried out manually in COOT (51) and PHENIX respectively (50) but failed to improve significantly the density and refinement statistics. A more complete model was obtained using the diffraction data from crystal 2. The structure was solved by molecular replacement with PHASER (52) using the SAD derived structure from crystal 1 as the search model. This vastly improved the resulting maps and statistics following subsequent rounds of manual model building and refinement. The final model of crystal 2 was refined to 1.97 Å resolution with Rwork/Rfree = 0.1927/0.2233 (Table 1) and contained four molecules in the asymmetric unit: molecule A, residues 1-121; molecule B, residues 1-131; molecule C, residues 1-121; molecule D, residues 1-131. Threonine

57 exists as a Ramachandran outlier with relatively weak density in the β3-β4 loop of molecules B and D respectively and could not be refined further. All structural models were rendered with Pymol (Schrödinger, Inc.) and surface electrostatics were calculated with APBS (53).

Size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS)

Purified HpR1Δ136 at 4mg/mL was subjected to size-exclusion chromatography using a Superdex 200 10/300 column (GE) equilibrated in SEC buffer (20mM HEPES pH 7.5, 150 mM KCl, 5 mM MgCl2, and 1mM DTT). The column was coupled to a static 18-angle light scattering detector (DAWN HELEOS-II) and a refractive index detector (Optilab T-rEX) (Wyatt Technology). Data were collected continuously at a flow rate of 0.5 mL/min. Data analysis was carried out using the program Astra VI. Monomeric BSA at 4mg/mL (Sigma) was used for normalization of the light scattering detectors and data quality control.

Preparation of oligonucleotide substrates

The following DNA oligonucleotides for filter binding synthesized commercially by Integrated DNA Technologies (IDT):

5mC_us 5’CCGGGTAAGA(5mC)CGGTAGCGAGCCCGG

5mC_ls 5’CCGGGCTCGCTA(5mC)CGGTCTTACCCGG

nm_us 5’CCGGGTAAGACCGGTAGCGAGCCCGG

nm_ls 5’CCGGGCTCGCTACGGTCTTACCCGG

Ll_us 5’CAGATCTGACGCTAGAGCTT

Ll_ls 5’AAGCTCTAGCGTCAGATCTG

Llscr_us 5’CAGATCTCGACGTAGAGCTT

Llscr_ls 5’AAGCTCTACGTCGAGATCTG

Lyophilized single-stranded oligonucleotides were resuspended to 1 mM in 10 mM Tris-HCl and 1 mM EDTA and stored at -20°C until needed. Single-stranded oligonucleotides were 5’ end-labeled with (γ32P)ATP using polynucleotide kinase (New England Biolabs) and then purified on a P-30 spin column (BioRad) to remove unincorporated label. Duplex substrates were prepared by heating equimolar concentrations of complementary strands (denoted with suffixes ‘us’ and ‘ls’ indicating upper and lower strands) to 95°C for 15 minutes followed by cooling to room temperature overnight and then purification on an S-300 spin column (GE) to remove single stranded DNA. Four duplex DNA substrates were prepared: a methylated EcMcrB-specific substrate, 5mC (5mC_us and 5mC_ls); a non-methylated EcMcrB-specific substrate, nm (nm_us and nm_ls), a site-specific substrate containing the *L.* lactis LlaJI.R1 binding site, Ll (Ll_us and Ll_ls), and a substrate with the *L.* lactis LlaJI.R1 binding site sequence scrambled as a control. Llscr (Llscr_us and Llscr_ls).

Filter binding assays

The standard buffer for the DNA binding assays contained 25 mM MES (pH 6.5), 2.0 mM MgCl2, 0.1 mM DTT, 0.01 mM EDTA, and 40 μg/mL BSA. Binding was performed with purified HpR1Δ136 (wildtype or mutants) or EcMcrB FL at 30°C for 10 min in a 30 μL reaction mixture containing 14.5 nM unlabeled DNA and 0.5 nM

labelled DNA. Samples were filtered through KOH-treated nitrocellulose filters (Whatman Protran BA 85, 0.45 μM) using a Hoefer FH225V filtration device for approximately 1 min. Filters were subsequently analyzed by scintillation counting on a 2910TR digital, liquid scintillation counter (PerkinElmer). All measured values represent the average of at least two independent experiments and were compared to a negative control to determine fraction bound.

Electrophoretic mobility shift assays (EMSA)

The standard buffer for the EMSAs contained 10 mM Tris-HCl pH8.0, 100 mM NaCl, 1 mM $MgCl_2$, and 1 mM DTT. Binding was performed with purified HpR1Δ136 (wildtype or mutants) at 25°C for 30 min in a 16 μL reaction mixture containing 10 ng/μL of N^6-methyladenine-free λ-phage DNA (New England Biolabs) digested with BamHI and NdeI (New England Biolabs) and purified via a NucleoSpin® Gel and PCR Clean-Up kit (Machery-Nagel). Following incubation, samples were analyzed by 0.7% agarose gel in 1x TAE at 4°C and 80V for 90 min. All gels were stained with SYBR® Green in 1x TAE overnight at 25°C (Thermo-Fisher Scientific) and visualized using a BioRad Gel Doc™ EZ imager system.

Acknowledgements: We thank April Lee for assistance with initial crystallization experiments, the Northeastern Collaborative Access Team (NE-CAT) beamline staff at the Advanced Photon Source (APS) for assistance with remote X-ray data collection, Dr. Holger Sondermann for critical reading of the manuscript and use of his SEC-MALS instrument, Dr. John O'Donnell for assistance with SEC-MALS data analysis, and Dr.

Eric Alani and Dr. Carol Manhart for guidance and assistance with filter binding experiments.

Conflict of interest: The authors declare that they have no conflicts of interest with the contents of this
article.

Author contributions: CJH and JSC designed the study and analyzed data. CJH purified and crystallized HpR1Δ136, collected X-ray diffraction data, solved the structure, and built the model. CJH cloned and purified all mutant constructs and conducted all biochemical assays. CJH and JSC carried out computational Modeling. CJH and JSC wrote the manuscript.

REFERENCES

1. Labrie, S. J., Samson, J. E., and Moineau, S. (2010) Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.* **8**, 317-327
2. Loenen, W. A., Dryden, D. T., Raleigh, E. A., Wilson, G. G., and Murray, N. E. (2014) Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Res.* **42**, 3-19
3. Loenen, W. A., Dryden, D. T., Raleigh, E. A., and Wilson, G. G. (2014) Type I restriction enzymes and their relatives. *Nucleic Acids Res.* **42**, 20-44
4. Pingoud, A., Fuxreiter, M., Pingoud, V., and Wende, W. (2005) Type II restriction endonucleases: structure and mechanism. *Cell Mol Life Sci.* **62**, 685-707

5. Pingoud, A., Wilson, G. G., and Wende, W. (2014) Type II restriction endonucleases--a historical perspective and more. *Nucleic Acids Res.* **42**, 7489-7527

6. Raghavendra, N. K., Bheemanaik, S., and Rao, D. N. (2012) Mechanistic insights into type III restriction enzymes. *Front Biosci.* **17**, 1094-1107

7. Meisel A., Bickle T. A., Krüger, D. H., and Schroeder, C. (1992) Type III restriction enzymes need two inversely oriented recognition sites for DNA cleavage. *Nature* **355**, 467-469

8. Loenen, W. A., and Raleigh, E. A. (2014) The other face of restriction: modification-dependent enzymes. *Nucleic Acids Res.* **42**, 56-69

9. Raleigh, E. A., and Wilson, G. (1986) Escherichia coli K-12 restricts DNA containing 5-methylcytosine. *Proc. Natl. Acad. Sci. USA* **83**, 9070–9074

10. Sutherland, E., Coe, L., and Raleigh, E.A. (1992) McrBC: a multisubunit GTP-dependent restriction endonuclease. *J. Mol. Biol.* **225**, 327-348

11. Krüger, T., Wild, C., and Noyer-Weidner, M. (1995) McrB: a prokaryotic protein specifically recognizing DNA containing modified cytosine residues. *EMBO J.* **14**, 2661-2669

12. Gast, F. U., Brinkmann, T., Pieper, U., Krüger, T., Noyer-Weidner, M., and Pingoud, A. (1997) The recognition of methylated DNA by the GTP-dependent restriction endonuclease McrBC resides in the N-terminal domain of McrB. *Biol. Chem.* **378**, 975-982

13. Pieper, U., Schweitzer, T., Groll, D. H., and Pingoud, A. (1999) Defining the location and function of domains of McrB by deletion mutagenesis. *Biol. Chem.* **380**, 1225-1230

14. Stewart, F. J., Panne, D., Bickle, T. A., and Raleigh, E. A. (2000) Methyl-specific DNA binding by McrBC, a modification-dependent restriction enzyme. *J. Mol. Biol.* **298**, 611-622

15. Mulligan, E. A., Hatchwell, E., McCorkle, S. R., and Dunn, J. J. (2010) Differential binding of Escherichia coli McrA protein to DNA sequences that contain the dinucleotide m5CpG. *Nucleic Acids Res.* **38**, 1997-2005

16. Panne, D., Müller, S. A., Wirtz, S., Engel, A., and Bickle, T. A. (2001) The McrBC restriction endonuclease assembles into a ring structure in the presence of G nucleotides. *EMBO J.* **20**, 3210-3217

17. Pieper, U., and Pingoud, A. (2002) A mutational analysis of the PD...D/EXK motif suggests that McrC harbors the catalytic center for DNA cleavage by the GTP-dependent restriction enzyme McrBC from Escherichia coli. *Biochemistry* **4**, 5236-4524

18. Panne, D., Raleigh, E. A., and Bickle, T. A. (1999) The McrBC endonuclease translocates DNA in a reaction dependent on GTP hydrolysis. *J. Mol. Biol.* **290**, 49-60

19. Pieper, U., Groll, D. H., Wünsch, S., Gast, F. U., Speck, C., Mücke, N., and Pingoud, A. (2002) The GTP-dependent restriction enzyme McrBC from Escherichia coli forms high-molecular mass complexes with DNA and produces a cleavage pattern with a characteristic 10-base pair repeat. *Biochemistry* **41**, 5245-5254

20. Cohen-Karni, D., Xu, D., Apone, L., Fomenkov, A., Sun, Z., Davis, P. J., Kinney, S. R., Yamada-Mabuchi, M., Xu, S. Y., Davis, T., Pradhan, S., Roberts, R. J., and Zheng, Y. (2011). The MspJI family of modification-dependent restriction endonucleases for epigenetic studies. *Proc. Natl. Acad. Sci. USA* **108**, 11040-11045

21. Borgaro, J. G., and Zhu, Z. (2013) Characterization of the 5-hydroxymethylcytosine-specific DNA restriction endonucleases. *Nucleic Acids Res.* **41**, 4198-4206

22. Bair, C., and Black, L.W. (2007) Exclusion of Glucosyl-Hydroxymethylcytosine DNA Containing Bacteriophages. *J. Mol. Biol.* **366**, 779–789

23. Sukackaite, R., Grazulis, S., Tamulaitis, G., and Siksnys, V. (2012) The recognition domain of the methyl-specific endonuclease McrBC flips out 5-methylcytosine. *Nucleic Acids Res.* **40**, 7552-7562

24. Horton, J. R., Wang, H., Mabuchi, M. Y., Zhang, X., Roberts, R. J., Zheng, Y., Wilson, G. G., and Cheng, X. (2014) Modification-dependent restriction endonuclease, MspJI, flips 5-methylcytosine out of the DNA helix. *Nucleic Acids Res.* **42**, 12092-12101

25. Kazrani, A. A., Kowalska, M., Czapinska, H., Bochtler, M. (2014) Crystal structure of the 5hmC specific endonuclease PvuRts1I. *Nucleic Acids Res.* **42**, 5929-5936

26. Horton, J. R., Borgaro, J. G., Griggs, R. M., Quimby, A., Guan, S., Zhang, X., Wilson, G.G., Zheng, Y., Zhu, Z., and Cheng, X. (2014) Structure of 5-hydroxymethylcytosine-specific restriction enzyme, AbaSI, in complex with DNA. *Nucleic Acids Res.* **42**, 7947-7959

27. Shao, C., Wang, C., and Zang, J. (2014) Structural basis for the substrate selectivity of PvuRts1I, a 5-hydroxymethylcytosine DNA restriction endonuclease. *Acta Crystallogr.* **D70**, 2477-2486

28. O'Driscoll, J., Glynn, F., Cahalane, O., O'Connell-Motherway, M., Fitzgerald, G. F., and Van Sinderen, D. (2004) Lactococcal plasmid pNP40 encodes a novel, temperature-sensitive restriction-modification system. *Appl. Environ. Microbiol.* **70**, 5546-5556

29. O'Driscoll, J., Heiter, D. F., Wilson, G. G., Fitzgerald, G. F., Roberts, R., and Van Sinderen, D. (2006). A genetic dissection of the LlaJI restriction cassette reveals insights on a novel bacteriophage resistance system. *BMC Microbiol.* **6**, 40-52

30. O'Driscoll, J., Fitzgerald, G. F., and van Sinderen, D. (2005) A dichotomous epigenetic mechanism governs expression of the LlaJI restriction/modification system. *Mol Microbiol.* **57**, 1532-1544

31. Yang, X., Xu, M., and Yang, S. T. (2016) Restriction modification system analysis and development of in vivo methylation for the transformation of Clostridium cellulovorans. *Appl. Microbiol. Biotechnol.* **100**, 2289-2299

32. Hendrickson, W. A. (2014) Anomalous diffraction in crystallographic phase evaluation. *Q. Rev. Biophys.* **47**, 49-93

33. Holm, L., and Rosenström, P. (2010). Dali server: conservation mapping in 3D. Nucleic Acids Res. *38*, W545-549.

34. Zhou, X. E., Wang, Y., Reuter, M., Mücke, M., Krüger, D. H., Meehan, E. J., Chen, L. (2004) Crystal structure of type IIE restriction endonuclease EcoRII reveals an autoinhibition mechanism by a novel effector-binding fold. *J. Mol. Biol.* **335**, 307-319

35. Grazulis, S., Manakova, E., Roessle, M., Bochtler, M., Tamulaitiene, G., Huber, R., and Siksnys, V. (2005). Structure of the metal-independent restriction enzyme BfiI reveals fusion of a specific DNA-binding domain with a nonspecific nuclease. *Proc. Natl. Acad. Sci. USA* **102**, 15797-15802

36. Golovenko, D., Manakova, E., Tamulaitiene, G., Grazulis, S., and Siksnys, V. (2009) Structural mechanisms for the 5'-CCWGG sequence recognition by the N- and C-terminal domains of EcoRII. *Nucleic Acids Res.* **37**, 6613-6624

37. Golovenko, D., Manakova, E., Zakrys, L., Zaremba, M., Sasnauskas, G., Gražulis, S., and Siksnys V. (2014) Structural insight into the specificity of the B3 DNA-binding domains provided by the co-crystal structure of the C-terminal fragment of BfiI restriction enzyme. *Nucleic Acids Res.* **42**, 4113-4122

38. Boer, D. R., Freire-Rios, A., van den Berg, W.A., Saaki, T., Manfield, I. W., Kepinski, S., López-Vidrieo, I., Franco-Zorrilla, J. M., de Vries, S. C., Solano, R., Weijers, D., and Coll, M. (2014) Structural basis for DNA binding specificity by the auxin-dependent ARF transcription factors. *Cell* **156**, 577-589

39. Yamasaki, K., Kigawa, T., Seki, M., Shinozaki, K., and Yokoyama, S. (2013) DNA-binding domains of plant-specific transcription factors: structure, function, and evolution. *Trends Plant Sci.* **18**, 267-276

40. Tamulaitiene, G., Silanskas, A., Grazulis, S., Zaremba, M., and Siksnys, V. (2014) Crystal structure of the R-protein of the multisubunit ATP-dependent restriction endonuclease NgoAVII. *Nucleic Acids Res.* **42**, 14022-14030

41. Yamasaki, K., Kigawa, T., Inoue, M., Tateno, M., Yamasaki, T., Yabuki, T., Aoki, M., Seki, E., Matsuda, T., Tomo, Y., Hayami, N., Terada, T., Shirouzu, M., Osanai, T., Tanaka, A., Seki, M., Shinozaki, K., and Yokoyama, S. (2004)

Solution structure of the B3 DNA binding domain of the Arabidopsis cold-responsive transcription factor RAV1. *Plant Cell* **16**, 3448-3459

42. Waltner, J. K., Peterson, F. C., Lytle, B. L., and Volkman, B. F. (2005) Structure of the B3 domain from Arabidopsis thaliana protein At1g16640. *Protein Sci.* **14**, 2478-2483
43. King, G. J., Chanson, A. H., McCallum, E. J., Ohme-Takagi, M., Byriel, K., Hill, J. M., Martin, J.L., and Mylne, J. S. (2013) The Arabidopsis B3 domain protein VERNALIZATION1 (VRN1) is involved in processes essential for development, with structural and mutational studies revealing its DNA-binding surface. *J. Biol. Chem.* **288**, 3198-3207
44. Sasnauskas, G., Tamulaitiene, G., Tamulaitis, G., Calyševa, J., Laime, M., Rimšeliene, R., Lubys, A., and Siksnys, V. (2017) UbaLAI is a monomeric Type IIE restriction enzyme. *Nucleic Acids Res.* **45**, 9583-9594
45. Swaminathan, K., Peterson, K., Jack, T. (2008) The plant B3 superfamily. *Trends Plant Sci.* **13**, 647-655
46. Deo, R. C., Groft, C. M., Rajashankar, K. R., and Burley, S.K. (2002) Recognition of the rotavirus mRNA 3' consensus by an asymmetric NSP3 homodimer. *Cell* **108**, 71-81
47. Kabsch, W. (2010) XDS. *Acta Crystallogr*. **D66**, 125-132
48. Evans, P. R. (2006) Scaling and assessment of data quality, *Acta Crystallogr.* **D62**, 72-82
49. Sheldrick, G. M. (2008). A short history of SHELX. *Acta Crystallogr*. **A64**, 112-122
50. Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P.H. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr.* **D66**, 213-221

51. Emsley, P., Lohkamp, B., Scott, W.G. and Cowtan, K. (2010). Features and development of Coot. *Acta. Crystallogr.* **D66**, 486-501

52. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007) Phaser crystallographic software. *J. Appl. Cryst.* **40**, 658-674

53. Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L. E., Brookes, D. H., Wilson, L., Chen, J., Liles, K., Chun, M., Li, P., Gohara, D. W., Dolinsky, T., Konecny, R., Koes, D. R., Nielsen, J. E., Head-Gordon, T., Geng, W., Krasny, R., Wei, G. W., Holst, M. J., McCammon, J. A., and Baker, N. A. (2018) Improvements to the APBS biomolecular solvation software suite. *Protein Sci.* **27**, 112-128

FOOTNOTES

This work was supported by Cornell University and the National Institutes of Health grant GM120242 (to J.S.C). This work is based upon research conducted at the NE-CAT beamlines (24-ID-C and 24-ID-E) under the General User Proposals GUP-51113 and GUP-41829 (PI: J.S.C). NE-CAT is funded by the National Institutes of Health program project grant P41 GM103403. The Pilatus 6M detector on 24-ID-C beam line is funded by a NIH-ORIP HEI grant S10 RR029205. This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. JSC is a Meinig Family Investigator in the Life Sciences.

This article contains Fig. S1.

The atomic coordinates and structure factors (code 6C5D) have been deposited in the Protein Data Bank (http://wwpdb.org/).

The abbreviations used are: RM, restriction modification; Mod, modification, Res, restriction; MDRS, modification-dependent restriction system; 5mC, 5-methylcytosine; 5hmC; 5-hydroxymethylcytosine; 5ghmC, 5-glucosylhydroxymethylcytosine; Ec, *Escherichia coli*; R^MC, methylated binding site where R is a purine and ^MC is a methylcytosine; Hp, *Helicobacter pylori*; HpR1Δ136, the N-terminal DNA binding domain of *Helicobacter pylori* LlaJI.R1 protein; At, *Arabidopsis thaliana*; SeMet; Selenomethionine; SAD, single-wavelength anomalous diffraction; SEC, size exclusion chromatography; MALS, multi-angle light scattering

Table 1. Data collection and refinement statistics for HpR1Δ136

Data collection		
Model	Crystal 1	Crystal 2 (PDB ID:
X-ray Source	NECAT 24ID-C	NECAT 24ID-C
Wavelength (Å)	0.9791	0.9791
Spacegroup	P1	P1
a, b, c (Å)	37.47, 44.39, 84.78	37.53, 43.77, 85.09
α, β, γ (°)	98.04, 94.37, 98.52	97.87, 93.86, 97.77
Resolution (Å)[a]	83.53 – 1.93 (2.08 – 1.93)	83.97 – 1.97 (2.10 – 1.97)
Mosaicity	0.20871°	0.11033°
No. measured reflections[a]	197952 (30000)	152937 (23180)
No. unique reflections[a]	53208 (15168)	41551 (12151)
Completeness (%)[a]	91.5 (91.9)	93.6 (90.1)
Multiplicity[a]	3.72 (1.98)	3.68 (1.91)
R_{meas}[a]	0.072 (0.395)	0.096 (0.624)
Mean I/σ_I[a]	11.7 (2.54)	9.0 (1.56)
$CC_{1/2}$[a]	0.998 (0.903)	0.997 (0.709)
Phasing		
Initial F.O.M.	0.548	
No. Se Sites	4	
Refinement		
R_{work}/R_{free}		0.1927/0.2233
RMSD		
Bond lengths (Å)		0.016
Bond angles (°)		1.530
Ramachandran plot		
Favored (%)		96.98
Allowed (%)		2.62
Outliers (%)		0.40
Average B-Factor		37.98
Clashscore		5.33
No. Atoms		
Macromolecule		4369
Solvent		245

[a] Denotes values for the highest resolution shell

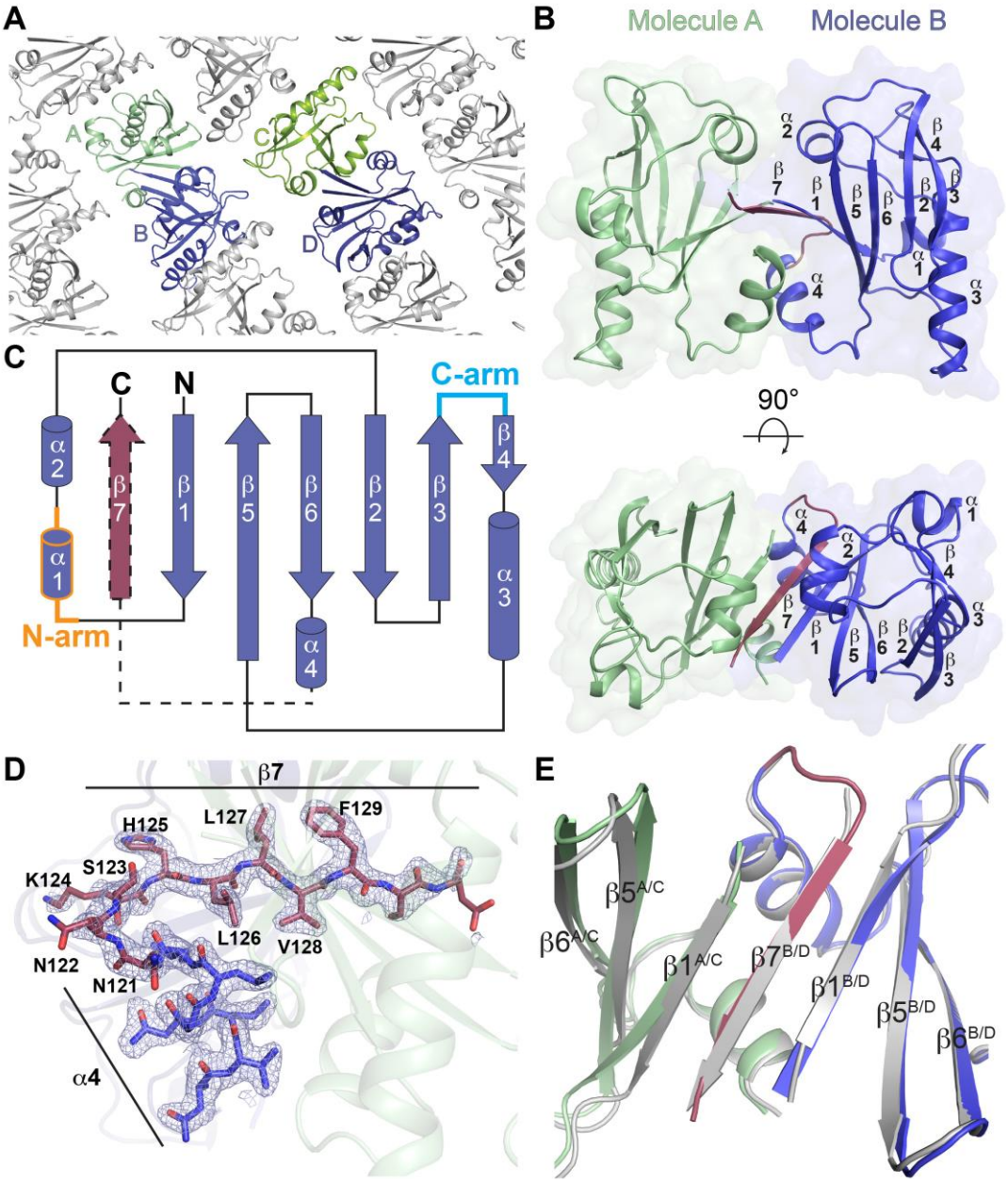

Figure 1. Structure and topology of HpR1Δ136. (A) Crystal packing of HpR1Δ136. AB and CD dimers are labeled. (B) Cartoon representations of HpR1Δ136 in two orientations. Molecules A and B are colored green and blue respectively. The asymmetric β7 strand is colored raspberry. (C) Topology diagram of HpR1Δ136. The core fold of each monomer is shown in blue. The relative position and connectivity of the asymmetric β7 strand associated with molecules B and D is denoted by dashed outlines and colored in raspberry. The N- and C-arms are colored orange and cyan respectively. (D) 2fo-fc electron density (blue mesh) of the α4-β7 region in molecule B contoured to 1σ. (E) Superposition of HpR1Δ136 dimers AB and CD. AB dimer colored green and blue while CD dimer colored gray. β7 from molecules B and D are colored raspberry and gray respectively.

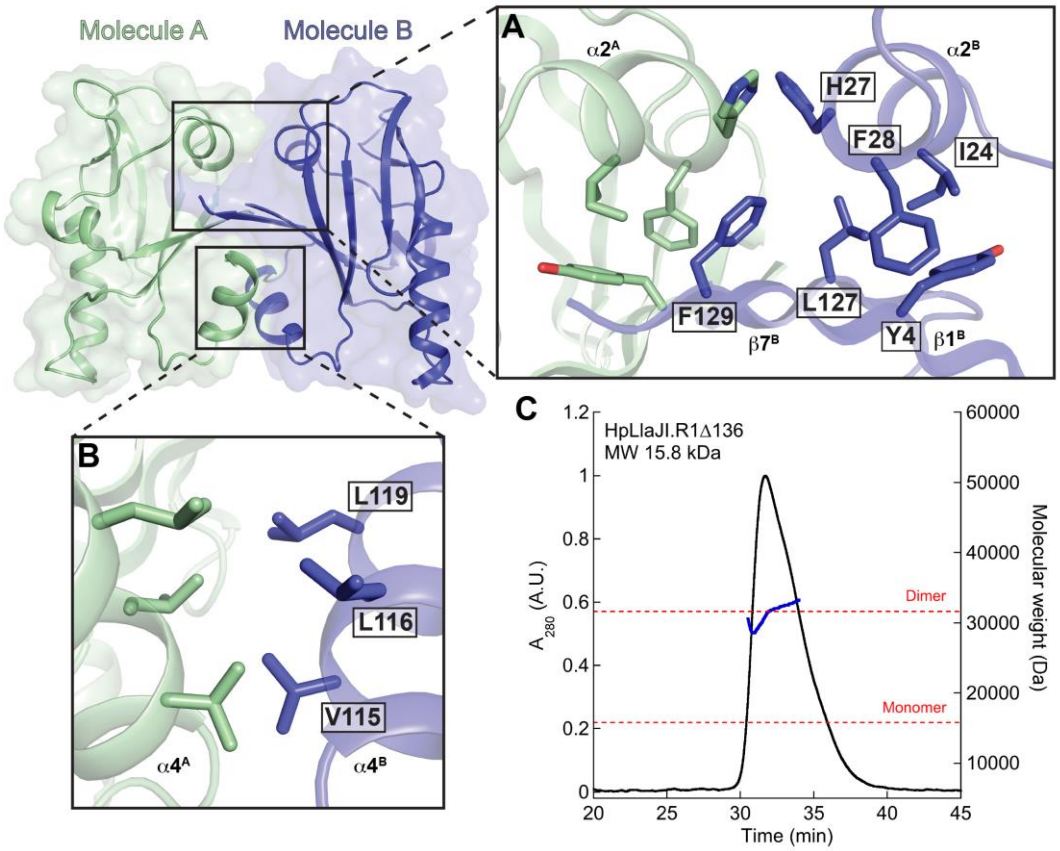

Figure 2. Critical structural contacts stabilizing the HpR1Δ136 dimer. (A) Zoomed view of β1, α2, and β7 interactions at the dimer interface. (B) Hydrophobic interactions between α4 helices across the dimer interface. (C) SEC-MALS analysis indicates HpR1Δ136 exists as a dimer in solution. UV trace (black) and calculated molecular weight based on light scattering (blue) are shown. Dashed red lines denote the predicted molecular weight of an HpR1Δ136 monomer and dimer.

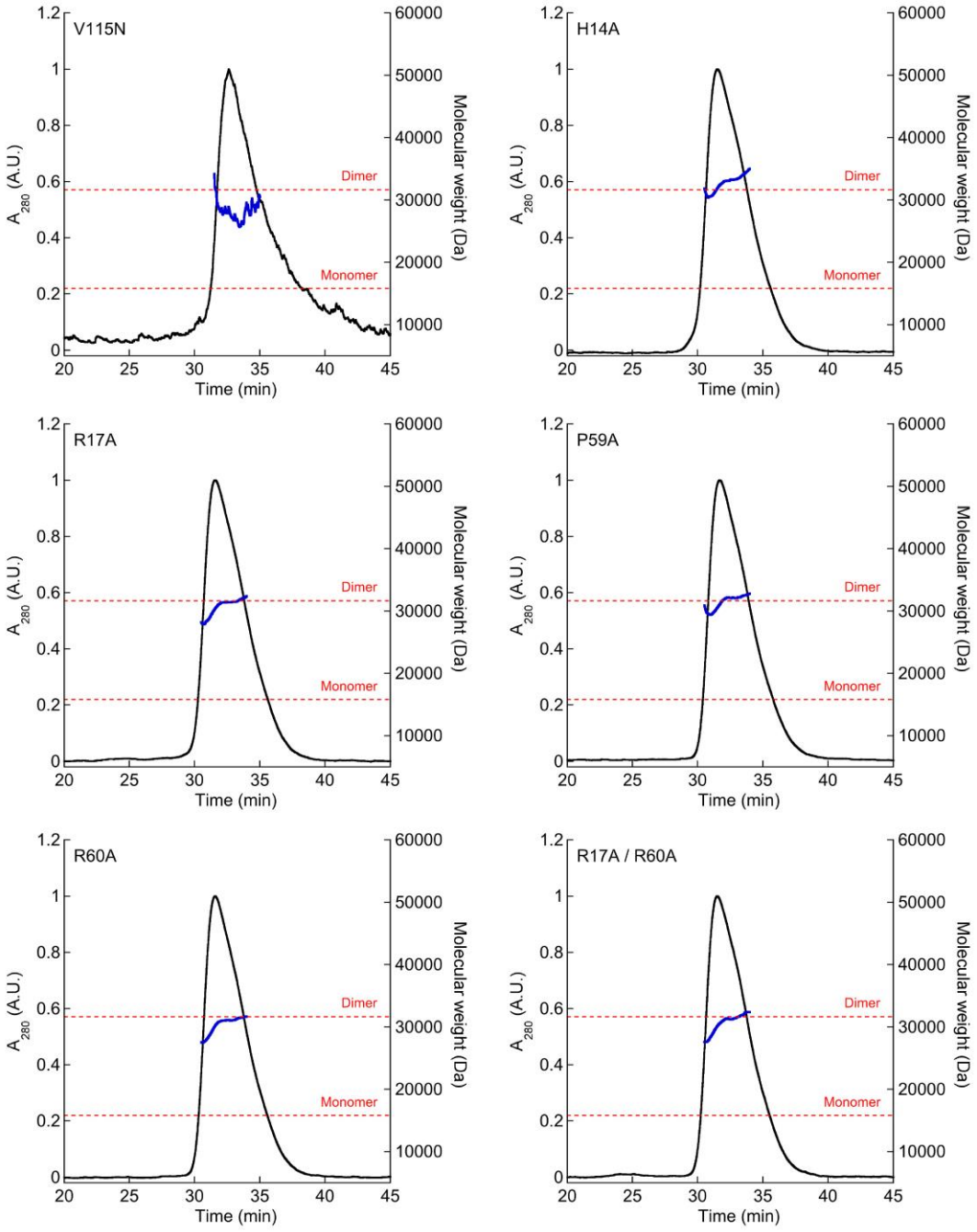

Figure 3. SEC-MALS of HpR1Δ136 mutants. V115N is located at the dimer interface (see Fig. 2B) while H14A, R17A, P59A, and R60A are putative binding site mutations based on structural homology (See Fig. 6D,E). UV trace (black) and calculated molecular weight based on light scattering (blue) are shown. Dashed red lines denote the predicted molecular weight of an HpR1Δ136 monomer and dimer.

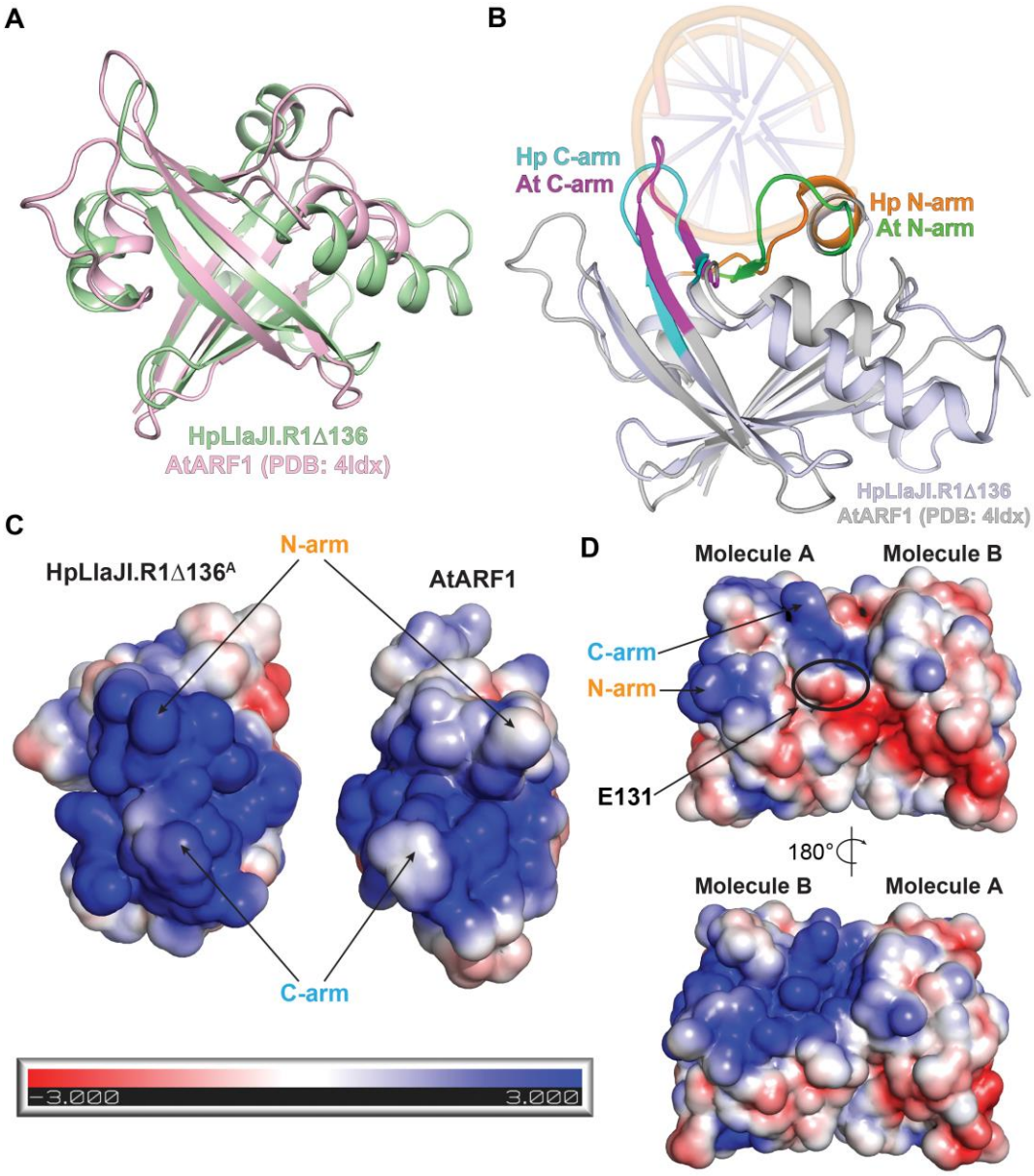

Figure 4. HpR1Δ136 adopts a B3 fold for site-specific DNA recognition. (A) Superposition of HpR1Δ136 (monomer A, green) with the *Arabidopsis thaliana* (At) ARF1 B3 domain (monomer A, pink; PDB: 4ldx) confirms similar structural fold. (B) Superposition of HpR1Δ136 (light blue) with the AtARF1A B3-DNA complex (gray; PDB: 4ldx) identifies the N-arm (orange and green) and C-arm (cyan and purple) regions that are necessary for target recognition. Structures are oriented looking down the helical axis of bound AtARF1 DNA. (C) Electrostatic surfaces of the individual HpR1Δ136 and AtARF1 B3 domains. Arrows indicate the N- and C-arms. Scale bar indicates electrostatic surface coloring from -3 K_bT/e_c to +3 K_bT/e_c. (D) Electrostatic surface of the HpR1Δ136 AB dimer in two orientations. N- and C-arms are labeled. The black circle indicates the location of E131 in molecule B.

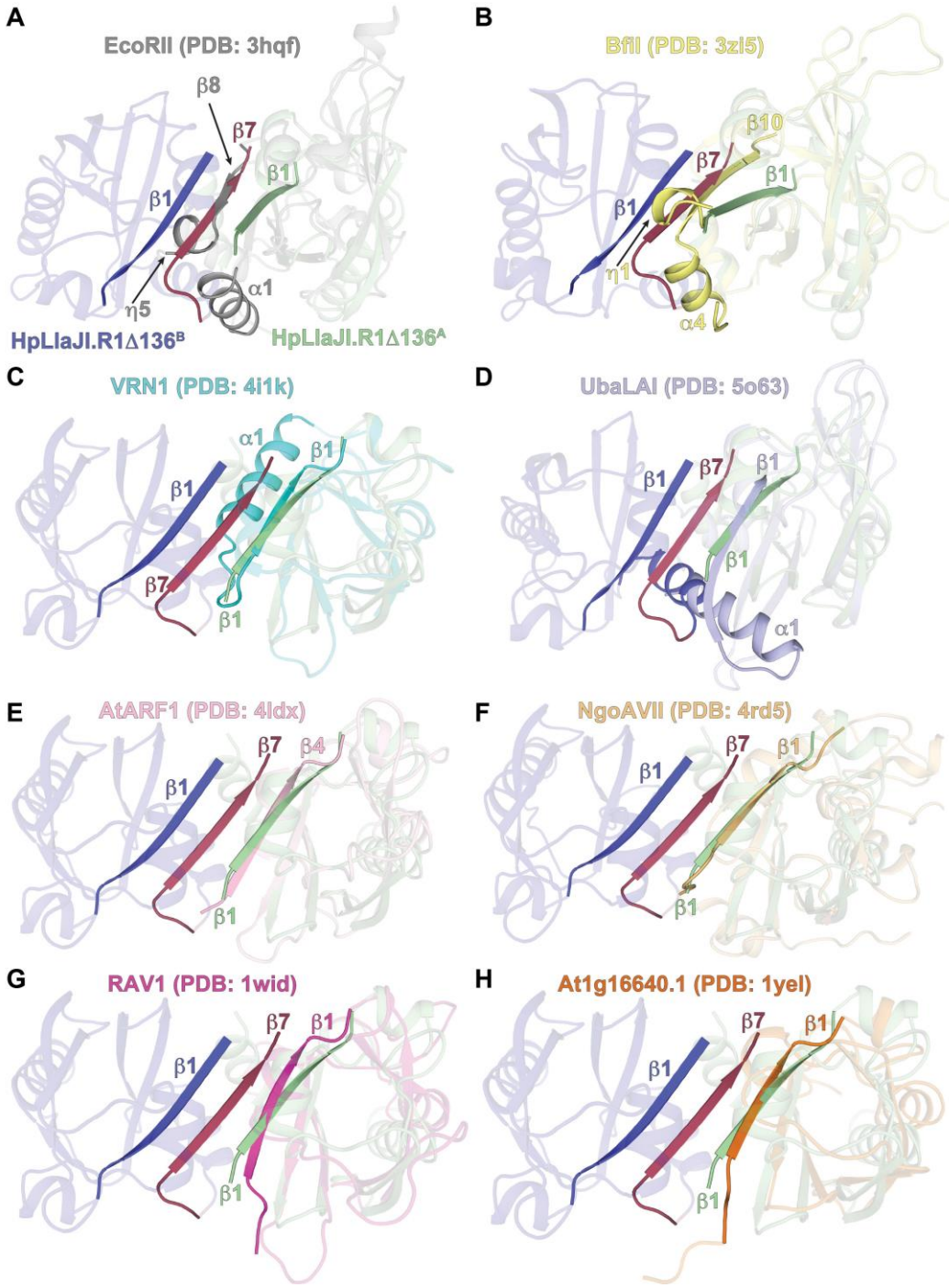

Figure 5. Structural constraints of B3 domain dimerization. Molecule A (green) and molecule B (blue) of HpR1Δ136 are shown in each panel with the asymmetric β7 strand (raspberry) and flanking β1 strands emphasized to delineate the dimer interface. Coordinates from different B3 domains were superimposed with molecule A and the overlapping structural elements that spatially align with the dimer interface are labeled and highlighted. PDB codes are indicated for each structure. (A) Superposition with EcoRII

(grey). (B) Superposition with BfiI (yellow). (C) Superposition with VRN1 (cyan). (D) Superposition with UbaLAI (light blue). (E) Superposition with AtARF1 (pink). (F) Superposition with NgoAVII (light orange). (G) Superposition with RAV1 (magenta). (H) Superposition with At1g16640.1 (orange).

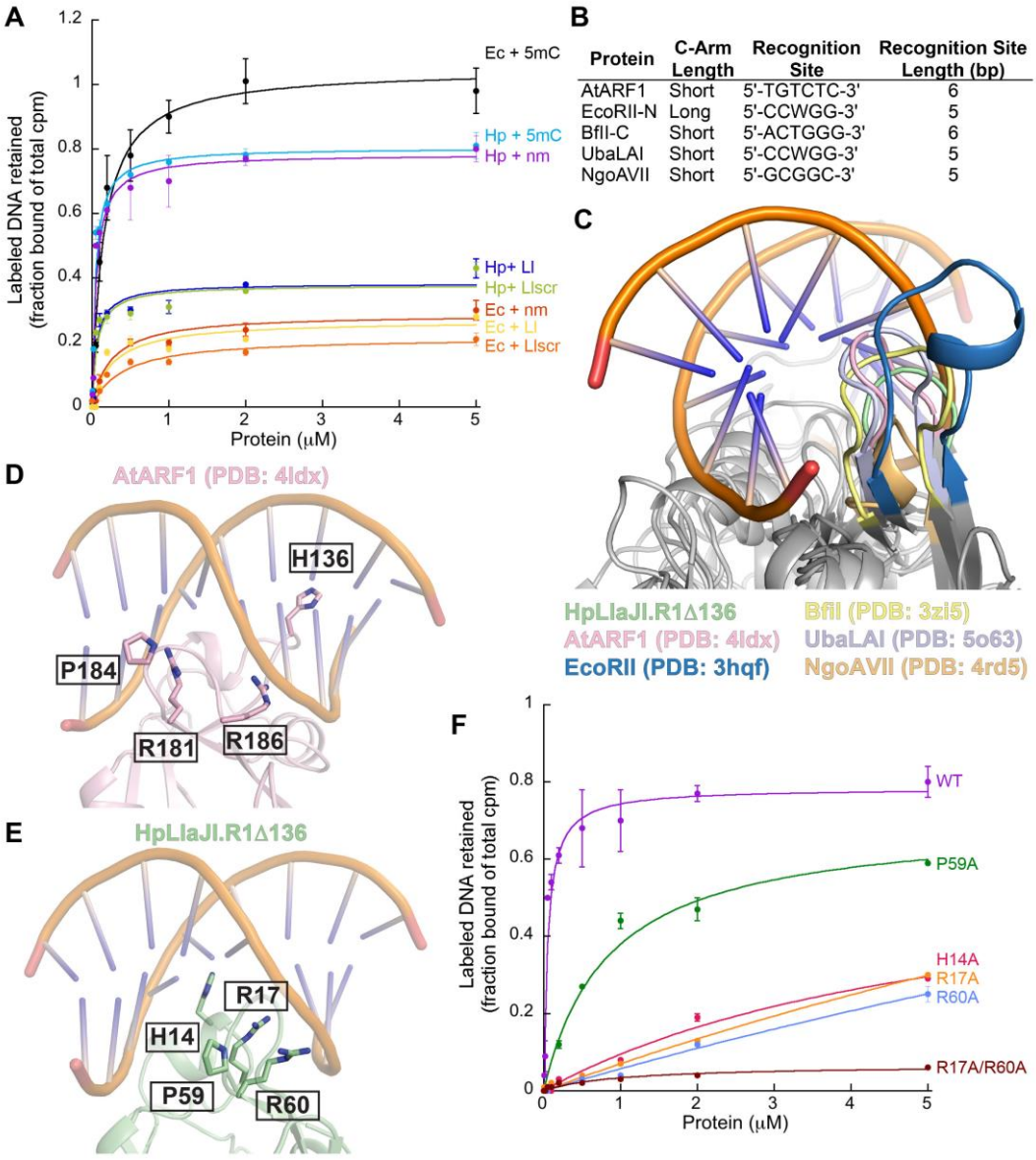

Figure 6. Structural modeling of HpR1Δ136 substrate recognition. (A) Filter binding analysis of HpR1Δ136 (Hp) and full-length EcMcrB (Ec) interactions with different DNA substrates. Substrates abbreviations are as follows: 5mC, methylated EcMcrB-specific substrate; nm, non-methylated EcMcrB-specific substrate; Ll, site-specific substrate containing the *L. lactis* LlaJI.R1 5'-GACGC-3'target site sequence; Llscr, substrate with the *L. lactis* LlaJI.R1 target site sequence scrambled as a control. Sequences for each substrate can be found in the Experimental Procedures. Binding was performed at 30°C for 10 min in a 30 μL reaction mixture containing 14.5 nM unlabeled DNA and 0.5 nM labelled DNA. Samples were filtered through KOH-treated nitrocellulose and binding was assessed by scintillation counting. (B) Relationship between C-arm length and target site length in previously determined B3 domain-DNA complexes. (C) Orientation of C-arm loops relative to DNA in various B3

domain homologs. DNA from the AtARF1 complex (PDB: 4ldx) is shown. C-arm coloring labeled below along with corresponding PDB codes. (D) Key residues in AtARF1 DNA binding. (E) Residues predicted to be important for HpR1Δ136 DNA binding based on structural comparison. AtARF1 DNA modeled as in (D). (F) Filter binding analysis of HpR1Δ136 mutants. Point mutations of predicted binding residues identified in (D) (H14A, red; R17A, orange; P59A, green; R60A, light blue; R17A/R60A; brown) were assessed for binding to the nm DNA substrate. Filter binding was carried out as described in (A). The wildtype curve (purple) is the same as shown in (A) (Hp+nm).

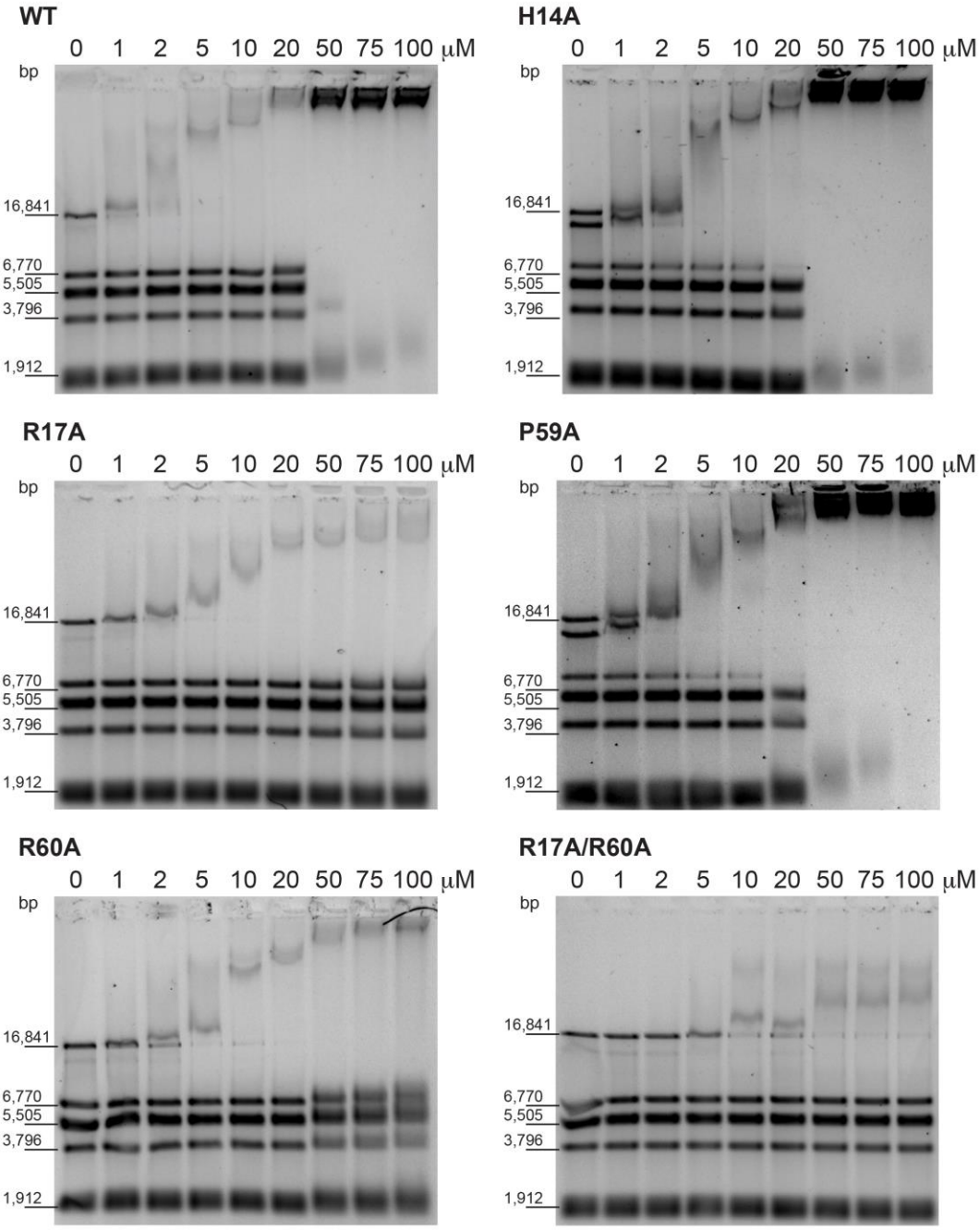

Figure 7. Electrophoretic mobility shift assay (EMSA) analysis of predicted HpR1Δ136 binding mutants. Binding was carried at 25°C for 30 min in a 16 μL reaction mixture containing 10 ng/μL of digested (BamHI/NdeI), non-methylated λ-phage DNA and increasing concentrations (0-100 μM) of each HpR1Δ136 construct. Gels were stained with SYBR® Green in 1x TAE overnight at 25°C. Calculated sizes (bp) of the digested DNA products are noted on the left of each gel.

Supporting information for:

The crystal structure of the *Helicobacter pylori* LlaJI.R1 N-terminal domain provides a model for site-specific DNA binding

Christopher J. Hosford[1] and Joshua S. Chappie[1,*]

From the [1]Department of Molecular Medicine, Cornell University, Ithaca NY 14853

*To whom correspondence should be addressed: Joshua S. Chappie: Department of Molecular Medicine, Cornell University, Ithaca NY 14853; chappie@cornell.edu; Tel. (607) 253-3654; Fax. (607) 253-3659.

Key words: LlaJI, restriction system, B3 domain, DNA binding, McrB, asymmetric dimerization

Content:

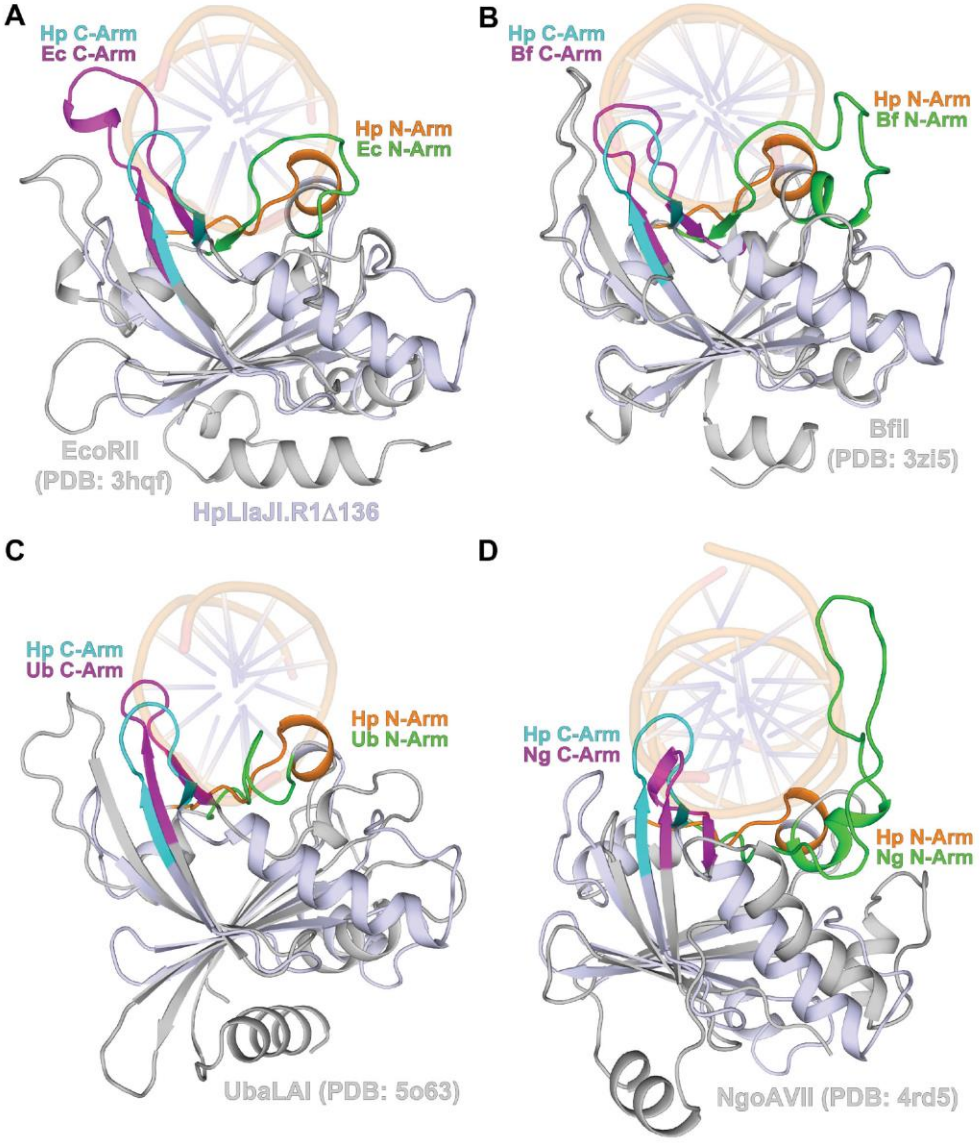

Figure S1. (Related to Figure 4). Structural superposition of HpR1Δ136 with other B3 domains supports model for DNA binding. Superposition of HpR1Δ136 (light blue) with different B3 domain-DNA complexes (gray). Structures are oriented looking down the helical axis of bound DNA. N- and C-arms are colored orange and cyan in HpR1Δ136 and green and purple in for all other structures. PDB codes are indicated for each model. (A) Superposition with the N-terminal region of the EcoRII restriction endonuclease. (B) Superposition with the C-terminal region of the BfiI restriction endonuclease. (C) Superposition with the N-terminal region of the UbaLAI restriction endonuclease. (D) Superposition with the C-terminal region of the NgoAVII restriction endonuclease.

Chapter 3. The crystal structure of the *Thermococcus gammatolerans* McrB N-terminal domain defines a new mode of substrate recognition and specificity in McrB homologs

The crystal structure of the *Thermococcus gammatolerans* McrB N-terminal domain defines a new mode of substrate recognition and specificity among McrB homologs

Christopher J. Hosford[1] and Joshua S. Chappie[1,*]

From the [1]Department of Molecular Medicine, Cornell University, Ithaca NY 14853

*To whom correspondence should be addressed: Joshua S. Chappie: Department of Molecular Medicine, Cornell University, Ithaca NY 14853; chappie@cornell.edu; Tel. (607) 253-3654; Fax. (607) 253-3659.

ABSTRACT

McrBC is a two-component, modification-dependent restriction system that cleaves foreign DNA containing methylated cytosines. Previous crystallographic studies show that *E. coli* McrB uses a base flipping mechanism to recognize these modified substrates with high affinity. The sidechains that facilitate DNA binding and stabilize the distorted duplex conformation are poorly conserved among McrB homologs, suggesting that other mechanisms may exist for binding modified DNA. Here we present the structures of the *Thermococcus gammatolerans* McrB DNA binding domain (TgΔ185) both alone and in complex with a methylated DNA substrate at 1.68Å and 2.27Å respectively. TgΔ185 consists of a YTH domain, which is commonly found in eukaryotic proteins that bind methylated RNA and is structurally unrelated to the *E. coli* McrB DNA binding domain. Structural superposition and co-crystallization identify a conserved aromatic cage in TgΔ185 that forms the binding pocket for a flipped-out base. Mutational analysis of this aromatic cage supports its role in conferring m^6A binding specificity. Surprisingly, TgΔ185 binds a single strand along the DNA duplex, which may have important functional implications. Together, these data underscore the idea that McrB

homologs have evolved different strategies for recognizing modified DNA and suggest the overall architecture of McrBC complex is modular, as its highly conserved motor and cleavage modules can be retargeted by attaching different DNA binding domains.

INTRODUCTION

Modification-dependent restriction systems (MDRs) recognize and cleave modified DNA (Labrie et al., 2010). Some enzymes like Mrr, McrA, MspJI, and McrBC are directed against methylated cytosines (Loenen and Raleigh, 2014) while others like GmrSD and members of the PvuRtsI family show specificity toward glucosylated nucleic acids (Bair and Black, 2007; Borgaro and Zhu, 2013). Collectively these proteins play a role in establishing the epigenetic landscape of bacterial genomes (Ishikawa et al., 2010) and are especially important in protecting against predatory bacteriophages, many of which incorporate modified bases into their DNA to evade detection by other defense systems (Weigele and Raleigh, 2016).

McrBC is a two-component, motor protein complex that was initially identified in *E. coli* (Ec) genetic screens by its ability to restrict glucosylation-deficient mutants of T4 phage (Luria and Human, 1952). EcMcrB is a 53 kDa protein with an N-terminal domain (pfam: DUF3578) that binds fully or hemi-methylated R^MC recognition elements (where R is a purine base and ^MC is a 4-methyl-, 5-methyl- or 5-hydroxymethyl-cytosine) (Sutherland et al., 1992; Krüger et al., 1995; Gast et al., 1997; Pieper et al., 1999b; Stewart et al., 2000, Zagorskaitė et al., 2018) and a C-terminal AAA+ (extended **A**TPases **A**ssociated with various cellular **A**ctivities) domain that binds/hydrolyzes GTP and mediates nucleotide-dependent oligomerization (Panne et

al., 2001). EcMcrB exhibits a low basal GTPase activity (~0.5-1 min-1) that can be stimulated ~30-40-fold via interaction with its partner EcMcrC (Pieper et al., 1999b), a 40 kDa protein that contains a C-terminal PD-(D/E)xK family endonuclease domain and lacks the ability to bind DNA on its own (Pieper and Pingoud, 2002). Biochemical studies suggest a model for cleavage in which EcMcrB and EcMcrC assemble at two R^MC sites separated by up to 3 kilobases and translocate DNA in a manner that depends on stimulated GTP hydrolysis (Panne et al., 1999). Collision of these assemblies cleaves both DNA strands near one of the R^MC sites (Stewart et al., 2000; Pieper et al., 2002), suggesting the complexes remain bound and translocate via DNA looping or twisting (Bourniquel and Bickle, 2002). These mechanochemical properties are reminiscent of Type I and Type III restriction-modification systems, which bind DNA at non-modified sites separated by up to thousands of base pairs and use ATP hydrolysis to power similar long-range translocation events that trigger cleavage either by collision or stalling (Dryden et al., 2001).

EcMcrB achieves specificity through a base flipping mechanism (Sukackaite et al., 2012, Zagorskaitė et al., 2018). Modified bases are rotated out of the DNA duplex and positioned into a pocket in the N-terminal domain, where they form numerous hydrogen bonds and hydrophobic interactions (Supplementary Figure 1). The concomitant insertion of a tyrosine residue (Y41) into the resulting gap stabilizes the duplex via base stacking. This strategy, while elegant, cannot simply be extrapolated to other McrB homologs as their N-terminal domains vary significantly in sequence, size, and predicted structural fold across different bacterial and archaeal species (Figure 1). In the handful of sequences that show identifiable homology to EcMcrB in this region

(e.g. Rhizobium sp. CF097), the tyrosine plug is not conserved and its mutation to the corresponding residue at that position – either alanine or glutamine – results in loss of DNA binding *in vitro* (Sukackaite et al., 2012). These findings imply that McrB homologs have evolved different mechanisms for substrate binding and/or may preferentially target other sequences and modifications. In support of this, we previously showed that the N-terminal domain of *Helicobacter pylori* LlaJI.R1 uses a B3 domain to recognize DNA site-specifically (Hosford and Chappie, 2018).

Here we present the crystal structures of the N-terminal DNA binding domain of *Thermococcus gammatolerans* McrB (TgΔ185) both alone and in complex with methylated DNA at 1.68Å and 2.27Å respectively. TgΔ185 is structurally distinct from the EcMcrB DNA binding domain, adopting a YTH domain fold commonly found in eukaryotic proteins that bind methylated RNA. Filter binding experiments show that TgΔ185 does not bind RNA and instead preferentially associates with 6-methyladenine-modified (m^6A) DNA. The TgΔ185-DNA complex indicates that different specificities is a result of an additive effect between the aromatic cage and differences in the electrostatic charge distribution compared to other YTH domains. Together these findings underscore the notion that McrBC is a modular nuclease that can be adapted to a broad array of targets.

RESULTS

TgMcrB does not preferentially bind m^5C DNA

To understand the broader species-specific determinants of McrB DNA binding, we purified both full-length *Thermococcus gammatolerans* (Tg) McrB and its isolated N-

terminal domain (TgΔ185, Figure 2A). Tg is a hyperthermophilic, radiation-tolerant archaea (Jolivet et al. 2003) and was selected for the enhanced thermostability of its proteins. Specificity for DNA containing methylated cytosines is a defining feature of EcMcrB (Sutherland et al., 1992; Krüger et al., 1995; Gast et al., 1997; Pieper et al., 1999b; Stewart et al., 2000, Zagorskaitė et al., 2018). As TgΔ185 share little sequence homology with the EcMcrB DNA binding domain (EcΔ155, Figure 2B), we first asked whether it could bind m5C modified DNA substrates. Initial characterization by analytical size exclusion chromatography (SEC) showed that TgΔ185 forms stable complexes similar to EcΔ155 (Figures 2C-D). To assess these interactions quantitatively, we examined the retention of radiolabeled m^5C and nonmethylated (nm) DNA in the presence of full-length TgMcrB or EcMcrB on alkaline-treated nitrocellulose filter paper (Papoulas, 1996). Filter binding shows that EcMcrB has a strong preference for m^5C DNA with a calculated binding constant on the order of ~200 nM (Figure 2E, Supplementary Table S3). TgMcrB, in contrast, binds both m^5C and nm DNA almost equally but with weaker affinity (calculated binding constants of ~1-2 μM) (Figure 2E, Supplementary Table S3). These data indicate TgMcrB is distinct from EcMcrB and displays a different sensitivity to modified DNA.

Tg∆185 adopts a YTH domain fold and preferentially binds m^6A DNA

To understand the molecular basis for the observed specificity differences, we determined the crystal structure of TgΔ185 at 1.68Å by selenium SAD phasing (Hendrickson, 2014) (Figure 3A). Tgδ185 is comprised of a six-stranded beta sheet – ordered β6-β1-β3-β4-β5-β2 – that is flanked by clusters of α-helices (Figure 3B). The

stands adopt a mainly antiparallel arrangement with only β1 and β3 oriented in a parallel fashion. The extended β4 strand subdivides the sheet and induces a sharp curvature that nearly folds the two opposing segments onto one another. Helical segments insert in loops that flank the beta sheet: α1 and α2 in the β1-β2 loop; α3 and α4 in the β4-β5 loop; α5 and α6 in the β5-β6 loop. Importantly, the overall topology of the Tgδ185 fold differs from that of Ecδ155 (Figure 3C,D).

The DALI alignment algorithm (Holm and Rosenström, 2010) indicates TgΔ185 shares structural homology with YTH domains (Z-score 7.5-8.5, RMSD 3.0-3.5) (Figure 3E,F). YTH domains are conserved RNA binding modules that specifically recognize N^6-methyladenosine (m^6A) modifications (Zhang et al., 2010, Liu et al., 2016). In eukaryotes, m^6A modifications are linked to the regulation of alternative splicing, RNA processing, mRNA degradation, and the circadian clock (Dominissini et al., 2012; Schwartz et al., 2013; Fustin et al., 2013, Wang et al., 2014). Given the structural similarity to YTH domains and lack of specificity toward m^5C DNA, we tested whether Tgδ185 can associate with m^6A modified RNA (Figure 4A). Filtering binding shows that while the human (Hs) YTHDC1 YTH domain specifically associates with m^6A RNA (calculated binding constant of 0.5931 μM), TgΔ185 shows little affinity for either methylated or non-methylated RNA substrates. We next asked whether TgΔ185 could bind m^6A modified DNA. Surprisingly, TgΔ185 associates more tightly with m^6A dsDNA, exhibiting a ~6.5-fold increase in affinity compared to m^5C or non-methylated dsDNA substrates (Figure 4B, Supplementary Table S3). This enhancement appears to be driven solely by the modification, as single stranded DNA oligos show the same binding profile (Figure 4C). These data indicate that TgΔ185 is a

DNA-specific YTH domain that preferentially targets substrates containing m^6A modifications.

An aromatic cage in TgΔ185 confers specificity for m^6A DNA

Crystallographic studies have shown that YTH domains recognize m^6A via a conserved "aromatic cage", wherein two to three aromatic residues provide stabilizing pi-stacking and hydrophobic interactions (Li et al., 2014; Luo and Tong, 2014; Theler et al., 2014; Zhu et al., 2014; Xu et al., 2014; Xu et al., 2015; Wang et al., 2016). Structural superposition with the m^6A-bound YTH domain from HsYTHDF2 (PDB ID: 4rdn, Z-score 8.5, RMSD 3.1) identifies W53, W115, and F121 as putative cage residues in TgΔ185, poised to serve as a binding site for modified bases (Figure Supplementary Figure S2).

To confirm this hypothesis, we determined the crystal structure of TgΔ185 in complex with DNA (Figure 5A). Although TgΔ185 crystallized with a variety of different modified substrates, suitable diffraction could only be obtained with a 19-mer dsDNA substrate that had single base pair overhangs and contained two mismatches flanking the internal m^5C modifications in each strand (m^5C dsDNA mm, Figure 5B). Initial maps at 2.64Å revealed defined DNA density associated with each TgΔ185 monomer and strong peaks for backbone phosphates. Numerous bases throughout the duplex, however, remained poorly resolved. An incomplete model for the Tgδ185-m5C dsDNA mm complex was built and used for molecular replacement into a 2.27Å resolution isomorphous data set. The higher resolution data set yielded vastly improved phases and interpretable electron density for both a stabilized base bound within the

aromatic cage and base pairs within the surrounding DNA duplex (Figure 5C-E).

The asymmetric unit contains a single TgΔ185 monomer bound to a six base pairs of DNA (Figure 5C, yellow). These DNA segments pack end-to-end, forming a pseudo-continuous duplex throughout the crystal lattice that is highly distorted (Figure 5F). The significant widening of the major groove (Figure 5F) likely arises from both TgΔ185-induced base flipping (Figure 5D) and the presence of mismatches in the DNA substrate that enhanced crystallization (Figure 5B). Tgδ185 decorates the extended duplex along a single strand (Figure 5C). This overall organization has two important implications. First, it suggests that TgΔ185 can associate with multiple sites along the 19-mer substrate. Second, it implies that resulting electron density attributed to the DNA does not reflect its unique sequence and instead represents an average distribution of the bases over the length of the duplex. We modelled the flipped-out bases as adenines since they repeat every six bases in the sequence of the crystallized substrate (Figure 5B). This yielded the best R_{free} value and strongest base density compared to refinement with alterative sequence registers. The apo- and DNA-bound Tgδ185 monomers superimpose with an average r.m.s. deviation of 0.549 Å, indicating no significant structural changes occur in the protein upon substrate binding.

To identify key structural features important in m^6A DNA binding, we compared Tgδ185-m^5C dsDNA mm to the HsYTHDC1 YTH domain complexed to m^6A-ssRNA (PDB ID: 4r3i, Z-score 7.7, RMSD 3.3). Like HsYTHDC1, Tgδ185 contains a large basic patch coincident with the aromatic cage to electrostatically interact with nucleic acids (Figure 6A,B). The binding and spatial orientation of m^6A in the aromatic cage of HsYTHDC1 (W377, W428, and I439) is nearly identical to that of Tgδ185 (Figure

6C,D). Although mutagenesis of either tryptophan residues in HsYTHDC1 perturbs m^6A RNA binding, mutating any of the aromatic cage residues in TgMcrB only reduces DNA binding by 7-fold and is comparable to the levels of m^5C or nm DNA binding (Figure 6E, Supplementary Table S3). No additional loss of binding was gained with a triple mutant to alanine.

Several arginine residues contribute to the basic patch in TgMcrBδ185 and are involved in DNA binding. Residues R55 and R168 form electrostatic interactions with the phosphate backbone on opposite strands. Although two arginine residues are shown to also engage with RNA in HsYTHDC1, only one is engaged with the phosphate backbone (R404) while the other stabilizes the resulting gap due to base flipping (R475) and □-stacks with the G-1 base. Mutagenesis of R475 in HsYTHDC1 to phenyalanine diminishes binding by 9-fold while mutating R475 to alanine decreases binding affinity over 100-fold (Xu et al., 2014). Two additional arginine residues in Tgδ185, R78 and R81, engage the DNA bases from the major groove (Supplementary Figure S3C). Although R81 could not be modelled due to insufficient electron density, their spatial similarity to R475 in HsYTHDC1 made us question their importance in DNA binding. However, mutating either R78 or R81 to alanine had no effect on DNA binding (Supplementary Figure S3D, Supplementary Table S3). Additional residues that make direct contact to DNA in the major groove (Y61 and N82) or form hydrogen bonds to the flipped-out adenine base (E16 and N19) also show no decrease in DNA binding (Supplementary Figure S3A,B, Supplementary Table S3). Interestingly, a double mutant of Y61 and N82 to alanine resulted in a 7-fold increase to DNA binding (Supplementary Figure S3D, Supplementary Table S3).

DISCUSSION

Here we described the crystal structures of Tgδ185 alone and in complex with DNA. TgΔ185 adopts a YTH domain fold and shows preference for m^6A DNA *in vitro*. This specificity distinguishes TgΔ185 from every other previously characterized YTH domain, all of which specifically target modified RNA. Canonical YTH domains use base flipping and an aromatic cage to recognize the m^6A modification (Li et al., 2014; Luo and Tong, 2014; Theler et al., 2014; Zhu et al., 2014; Xu et al., 2014; Xu et al., 2015; Wang et al., 2016). Our structural data indicate that TgΔ185 employs the same general strategy. Mutation of aromatic cage residues in other YTH domains completely abolishes RNA binding (Xu et al., 2014). Surprisingly, the W53A/W115A/F121A triple mutant only reduces TgΔ185's binding to m^6A DNA by ~7-fold, equivalent to its affinity for non-methylated DNA. This argues that the aromatic cage alone dictates the specificity for the m^6A modification while other structural features contribute toTgΔ185's overall DNA binding.

The overall electrostatic surface area of Tgδ185 varies significantly from HsYTHDC1. This is largely due to the varying number of arg residues on the exterior surface surrounding the aromatic cage. In the YTH domains that specifically bind RNA, such as HsYTHDC1 and HsYTHDF2, only two arg residues are within proximity to bind RNA (R404/R475 in HsYTHDC1 and R441/R527 in YTHDF2) (Xu et al., 2014, Li et al., 2014). In contrast, Tgδ185 contains four arg residues in proximity to the bound DNA (R55, R78, R81, and R162). Of these, R78 and R81 appear to form base specific contacts within the major groove while R55 and R162 form electrostatic, 'clamp' interactions with the phosphate backbone in opposite strands. We hypothesize that this

even distribution of arg residues is a primary driver of the specificity of TgΔ185 for DNA over RNA.

Additional sequence specific interactions have been identified in other YTH domains that are structurally conserved in TgΔ185 (Xu et al., 2014, Li et al., 2014). In the YTHDF1 YTH domain structure (PDB ID: 4rcj), residue Y397 wedges between the G-1 and m^6A bases and confers a preference for binding the sequence Gm^6A in RNA. The Y61 residue in TgΔ185 is spatially orchestrated similarly to Y397 in YTHDF1 and forms a wedge in the major groove with N82. Interestingly, mutagenesis of Y61 and N82 to alanine increases TgΔ185s affinity for m^6A DNA by nearly 7-fold (Supplementary Figure S2, Supplementary Table S3). We hypothesize this may be a result of a sequence specific contact that would otherwise be made with an ideal substrate. Mutagenesis of Y61 and N82 may therefore increase TgΔ185 binding tolerance for. Additional screening of different bases preceding the m^6A base (specifically at positions -1 and -2) is necessary to validate this observation.

Canonical McrB as defined in the *E. coli* homolog has been shown to preferentially bind m^5C DNA. Therefore, the preference of TgMcrB for m^6A over m^5C was unexpected. We recently published the crystal structure of the N-terminal domain of the *Helicobacter pylori* LlaJI.R1, a site-specific McrB homolog, that revealed it recognizes DNA site-specifically via a B3 domain (Hosford and Chappie, 2018). These structures reveal different specificities and modes of substrate recognition amongst the N-terminal domains of McrB homologs. Interestingly the *E. coli* McrB has been previously shown to contain a cryptic translational start site in the *mcrB* gene that encodes for a shorter version of McrB. This shorter length peptide includes only the

motor domain and is thought to play a regulatory role in inhibiting McrBC activity (Dila D, 1990). Its GTPase activity is fully functional and can be stimulated by McrC but cannot bind or cleave DNA. Taken together, it appears that McrB is a modular restriction system that only needs an N-terminal domain to target it to DNA and can be used as a platform for the engineering of novel restriction enzymes.

Because of its specificity for m^5C, EcMcrBC is commonly used as a diagnostic tool to monitor epigenetic changes underlying mammalian gene expression (Fouse et al., 2010), tissue specific development (Santoso et al., 2000), and perturbation to normal methylation patterning associated with human diseases like Prader-Willi and Angelman syndromes (Chotai and Payne, 1998) and Fragile-X mental retardation (Burman et al., 1999). Recent studies have implicated N6-methyladenine modification as an important epigenetic maker in mammalian cells (Luo and He, 2017, NSMB; Xiao et al. 2018). Our structural and biochemical results suggest TgMcrBC could be utilized in a similar capacity to track m^6A methylation.

EXPERIMENTAL PROCEDURES

Identification and phylogenetic analysis of McrB homologs

Putative McrB homologs were initially identified by BLAST using the sequence of the *E. coli* McrB AAA+ domain to search against the DOE Integrated Microbial Genomes (IMG/ER) database (Chen et al., 2017). These candidates were only considered if they contained the conserved McrB consensus motif MNxxDRS and the presence of an adjacent McrC gene could be confirmed by neighbor analysis. Homologs were then subdivided into groups according to their divergent N-terminal domains. A phylogenetic

tree incorporating a representative from each group was generated using the DOE IMG/ER analysis tools. Structural fold prediction for each unique N-terminal domain was carried out using the Phyre 2 protein fold recognition server (Kelley et al., 2015).

Cloning, expression, and purification of TgMcrB constructs

DNA encoding the *T. gammatolerans EJ3* McrB protein (DOE IMG ID 644807740) was codon optimized for *E. coli* expression and synthesized commercially by GENEART. DNA encoding full-length TgMcrB was amplified by PCR and cloned into pET21b, introducing a 6xHis tag at the C-terminus. DNA encoding the N-terminal domain (TgΔ185, residues 1-185) was amplified by PCR and cloned into pET15bP, a modified pET15b (Novagen) plasmid in which an Hrv3C protease site (LEVLFQGP) replaces the thrombin site after the N-terminal 6xHis tag. Native TgMcrB and TgΔ185 were transformed into BL21(DE3) cells, grown at 37°C in Terrific Broth to an A_{600} of 1.0, and then induced with 0.3 mM isopropyl 1-thio-β-D-galactopyranoside (IPTG) overnight at 19°C. All cells were harvested, washed with nickel load buffer (20 mM HEPES pH 7.5, 500 mM NaCl, 30 mM imidazole, 5% glycerol (v/v), and 5 mM β-mercaptoethanol), and pelleted a second time. Pellets were flash frozen in liquid nitrogen and stored at -80°C. Seleno-methionine labeled (SeMet) TgΔ185 was expressed in minimal media in the absence of auxotrophs as described previously (Van Duyne et al., 1993). Thawed Pellets from 500-mL cultures were resuspended in 30-ml of nickel load buffer supplemented with 10 mM phenylmethylsulfonyl fluoride (PMSF), 5 mg of DNAse (Roche), 5 mM $MgCl_2$, and a complete protease inhibitor cocktail tablet (Roche). Lysozyme was added to 1 mg/ml and the mixture was incubated for 15 min

rocking at 4°C. Cells were disrupted by sonication and the lysate was cleared of debris by centrifugation at 13,000 rpm (19,685 × g) for 30 minutes at 4°C. For native and SeMet TgΔ185, the supernatant was filtered, loaded onto a 5-ml HiTrap chelating column charged with $NiSO_4$ and then washed with nickel load buffer. TgΔ185 was eluted with an imidazole gradient from 30mM to 1M. Pooled fractions were dialyzed overnight at 4°C into Ni loading buffer with reduced salt (50mM NaCl) in the presence of Hrv3C protease to remove the N-terminal His tag. The sample was reapplied to a 5-ml HiTrap chelating column charged with $NiSO_4$. The flow through was fractionated to collect cleaved TgΔ185, concentrated, and further purified by size exclusion chromatography (SEC) using a Superdex 75 16/600 pg column. For full-length TgMcrB, the supernatant from sonication was filtered, heated to 65°C for 20 min, centrifuged at 4,000 rpm (6,057 × g) for 10 min at 4°C, and filtered again prior to purification on a 5-ml HiTrap chelating column as described above. Pooled peak fractions were concentrated and purified further by SEC. All proteins were exchanged into a final buffer of 20mM HEPES pH 7.5, 150mM KCl, 5 mM MgCl2, and 1mM DTT during SEC and concentrated to 5-40 mg/ml. SeMet TgΔ185 was purified similarly but was supplemented with 5 mM DTT in the SEC buffer. TgMcrB mutants were generated by Quikchange mutagenesis (Agilent Technologies) and confirmed by sequencing.

Cloning, expression, and purification of EcMcrB Δ155

DNA encoding the full-length *E. coli* McrB protein (Uniprot P15005; DOE IMG ID 646316336) was codon optimized for *E. coli* expression and synthesized commercially by GENEART. DNA encoding the N-terminal domain (EcΔ155, residues 1-155) was

cloned into pMAL-c2Xp, a modified pMAL-c2X (New England Biolabs) plasmid in which an Hrv3C protease site replaces the Factor Xa site after the N-terminal MBP tag. Ecδ155 was transformed into BL21(DE3) cells, grown at 37°C in Terrific Broth to an A_{600} of 1.0, and then induced with 0.3 mM IPTG overnight at 19°C. All cells were harvested, washed with TGED500 (20 mM Tris-HCl pH 8.0, 500 mM NaCl, 1 mM EDTA, 5% glycerol (v/v), and 1 mM DTT), and pelleted a second time. Pellets were flash frozen in liquid nitrogen and stored at -80°C. Thawed Pellets from 500-mL cultures were resuspended in 30-ml of TGED500 supplemented with 10 mM PMSF, 5 mg of DNAse (Roche), 5 mM $MgCl_2$, and a complete protease inhibitor cocktail tablet (Roche). Lysozyme was added to 1 mg/ml and the mixture was incubated for 15 min rocking at 4°C. Cells were disrupted by sonication and the lysate was cleared of debris by centrifugation at 13,000 rpm (19,685 × g) for 30 minutes at 4°C. The supernatant was filtered, loaded onto 30-40 ml of amylose resin, washed with TGED500, and eluted with TGED500 supplemented with 10 mM D-maltose. Pooled fractions were dialyzed overnight at 4°C into TGED with reduced salt (TGED50, 50mM NaCl) in the presence of Hrv3C protease to remove the N-terminal MBP tag. The sample was then applied to a 5-ml HiTrap Q HP ion exchange column in TGED50 and eluted with a NaCl gradient from 50mM to 500mM. Pooled fractions were concentrated and further purified by SEC using a Superdex 75 10/300 GL column. Ecδ155 was exchanged into a final buffer of 20mM HEPES pH 7.5, 150mM KCl, 5 mM MgCl2, and 1mM DTT during SEC and concentrated to 5-40 mg/ml.

Cloning, expression, and purification of HsYTHDC1

DNA encoding the *Homo sapiens* (Hs)YTHDC1 YTH domain (residues 344-509) was codon optimized for *E. coli* expression and synthesized commercially by Integrated DNA Technologies (IDT) and cloned into pET15bP. The HsYTHDC1 344-509 was transformed into BL21(DE3) cells, grown at 37°C in Terrific Broth to an A_{600} of 1.0, and then induced with 0.3 mM IPTG overnight at 19°C. All cells were harvested, washed with nickel load buffer, and pelleted a second time. Pellets were flash frozen in liquid nitrogen and stored at -80°C. Thawed Pellets from 500-mL cultures were resuspended in 30-ml of nickel load buffer supplemented with 10 mM PMSF, 5 mg of DNAse (Roche), 5 mM $MgCl_2$, and a complete protease inhibitor cocktail tablet (Roche). Lysozyme was added to 1 mg/ml and the mixture was incubated for 15 min rocking at 4°C. Cells were disrupted by sonication and the lysate was cleared of debris by centrifugation at 13,000 rpm (19,685 × g) for 30 minutes at 4°C. The supernatant was filtered, loaded onto a 5-ml HiTrap chelating column charged with $NiSO_4$, washed with nickel load buffer, and eluted with an imidazole gradient from 30mM to 1M. Pooled fractions were concentrated and further purified by SEC using a Superdex 75 10/300 GL column. HsYTHDC1 344-509 was exchanged into a final buffer of 20mM HEPES pH 7.5, 150mM KCl, 5 mM MgCl2, and 1mM DTT during SEC and concentrated to 5-40 mg/ml.

Preparation of oligonucleotide substrates

All DNA and RNA substrates for analytical SEC, filter binding, and crystallization were purchased from IDT. Lyophilized non-methylated and HPLC-purified modified single-

stranded oligonucleotides were resuspended in to 1 mM in 10 mM Tris-HCl and 1 mM EDTA and stored at -20°C until needed. Single-stranded oligonucleotides were 5′ end-labeled with (γ-^{32}P)ATP using polynucleotide kinase (New England Biolabs) and then purified on a P-30 spin column (Bio-Rad) to remove unincorporated label. Duplex substrates were prepared by heating equimolar concentrations of complementary strands (denoted with suffixes "us" and "ls" indicating upper and lower strands) to 95 °C for 15 min followed by cooling to room temperature overnight and then purification on an S-300 spin column (GE Healthcare) to remove ssDNA. Supplementary Table S1 shows the sequence of each oligonucleotide used in this work.

Analytical size exclusion chromatography (SEC)

Samples (50 μl) of 100 μM EcMcrB δ155 or TgMcrB δ185 were mixed with m^5C dsDNA in a 2:1.2 molar ratio in 20 mM HEPES pH7.5, 150 mM KCl, 5 mM $MgCl_2$, and 1 mM DTT and incubated at room temperature for 10-15 min. Each reaction was fractionated via gel filtration on a Superdex 75 3.2/300 analytical SEC column equilibrated with 20 mM HEPES pH7.5, 150 mM KCl, 5 mM $MgCl_2$, and 1 mM DTT. Fractions containing samples were subjected to 4-20% gradient SDS-PAGE, silver stained for DNA, and Coomassie stained for protein.

Filter binding assays

The standard buffer for the DNA-binding assays contained 25 mM MES, pH 6.5, 2.0 mM$MgCl_2$, 0.1 mM DTT, 0.01 mM EDTA, and 40 μg/ml BSA. Binding was performed with purified TgMcrB FL (WT or mutants) or HsYTHDC1 3434-509 at 30°C for 10

min in a 30-μl reaction mixture containing 14.5 nM unlabeled DNA and 0.5 nM labeled DNA. Samples were filtered through KOH-treated nitrocellulose filters (Whatman Protran BA 85, 0.45 μM) using a Hoefer FH225V filtration device for ~1 min. Filters were subsequently analyzed by scintillation counting on a 2910TR digital, liquid scintillation counter (PerkinElmer Life Sciences). All measured values represent the average of at least three independent experiments and were compared with a negative control to determine fraction bound.

Crystallization, X-ray data collection, and structure determination

SeMet TgΔ185 was crystallized by sitting drop vapor diffusion in 0.1M MES pH6.5, 3.2M Amm. sulfate with a drop size of 2 μL and reservoir volume of 650 μl. Crystals appeared within 6-8 days at 20°C and were of the space group C2 with unit cell dimensions a = 67.84, b = 43.99, c = 61.96 and $\alpha = 90.00$, $\beta = 120.28$, $\gamma = 90.00$. Samples were cryoprotected with Parabar 10312 from Hampton Research and frozen in liquid nitrogen. Crystals were screened and optimized at the MacCHESS F1 beamline at Cornell University and single-wavelength anomalous diffraction (SAD) data were collected remotely on the tuneable NE-CAT 24-ID-C beamline at the Advanced Photon Source at the selenium edge energy at 12.663 keV (0.9791 Å) (Table S2). Data were integrated and scaled using the NE-CAT RAPD pipeline. Heavy atom sites were located using SHELX (Sheldrick, 2008) and phasing, density modification, and initial model building was carried out using the Autobuild routines of the PHENIX package (Adams et al., 2010). Further model building and refinement was carried out manually in COOT (Emsley et al., 2010) and PHENIX respectively (Adams et al., 2010). The final model

contained one molecule in the asymmetric unit containing residues 1-175 and was refined to 1.68Å resolution with R_{work}/R_{free} values of 0.1770/0.1907 (Supplementary Table S2).

SeMet TgΔ185 was crystallized in complex with a 19-mer DNA substrate containing a single m^5C modification in each strand (meC15 mismatched, Supplementary Table S2) by sitting drop vapor diffusion in 0.1M HEPES pH7.5, 20% PEG 3350, and 0.20M ammonium sulfate with a drop size of 2 μL and reservoir volume of 650 μl. MeC15 mismatched contained base pair mismatches flanking the m^5C sites, which were necessary to obtain diffraction quality crystals. TgΔ185 and meC15 mismatched DNA were mixed at a molar ratio of 2:1.2 and incubated at room temperature for 10-15 minutes prior to crystallization experiments. Crystals appeared within 10-14 days at 20°C and were of the space group $P2_12_12_1$ with unit cell dimensions a = 41.87, b = 56.50, c = 109.28 and α = 90.00, β = 90.00, γ = 90.00. Samples were cryoprotected with Parabar 10312 and frozen in liquid nitrogen. An initial 2.64 Å dataset (TgMcrB D185 + meC15 mismatched 1) was collected at NE-CAT 24-ID-E beamline at the selenium edge energy at 12.663 keV (0.9791 Å) and solved by molecular replacement in PHASER (McCoy et al., 2007) using the unbound TgΔ185 monomer structure determined from Se SAD phasing as the search model. A more complete model was obtained using the diffraction data from a second crystal, TgMcrB D185 + meC15 mismatched 2. The structure was solved by molecular replacement with PHASER (McCoy et al., 2007) using the MR-derived structure from TgMcrB D185 + meC15 mismatched 1 as the search model. This vastly improved the resulting maps and statistics following subsequent rounds of manual model building and refinement. The

final model of crystal 2 contained one molecule in the asymmetric unit containing residues 3-175 with 6 bp of the DNA substrate and was refined to 2.27 Å resolution with R_{work}/R_{free} = 0.2455/0.2851 (Supplementary Table S2).

Structural superpositions were carried out in Chimera (Pettersen et al., 2004). All structural renderings were generated using Pymol (Schrodinger) and surface electrostatics were calculated using APBS (Jurrus et al., 2018).

Acknowledgements

We thank Dr. Eric Alani and Carol Manhart for advice and assistance with filter binding assays, Drs. Eric Alani and Chris Fromme for critical reading of the manuscript, and the Northeastern Collaborative Access Team (NE-CAT) beamline staff at the Advanced Photon Source (APS) for assistance with remote X-ray data collection. We additionally thank Dr. Frederick Dyda for the generous use of a rotating anode home source for preliminary X-ray diffraction studies and cryo-protection optimization. The atomic coordinates of TgΔ185 and the TgΔ185-5mC DNA complex were deposited in the Protein Data Bank with accession numbers XXXX and XXXX respectively.

Funding

J.S.C. is a Meinig Family Investigator in the Life Sciences. This work is based upon research conducted at the Macromolecular Diffraction facility at the Cornell High Energy Synchrotron Source (MacCHESS) and the NE-CAT beamlines (24-ID-C and 24-ID-E). CHESS is supported by an award from the National Science Foundation (DMR-1332208) and MacCHESS is supported by a grant from the National Institutes

of Health (GM-103485). NE-CAT is funded by the National Institutes of Health program project grant (P41 GM103403). The Pilatus 6M detector on 24-ID-C beam line is funded by a NIH-ORIP HEI grant (S10 RR029205). This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. Funding for open access charge: Cornell University, laboratory startup funds.

REFERENCES

Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr. 2010 Feb;66(Pt 2):213-21.

Bair CL, Black LW. A type IV modification dependent restriction nuclease that targets glucosylated hydroxymethyl cytosine modified DNAs. J Mol Biol. 2007 Feb 23;366(3):768-78. Epub 2006 Nov 21.

Borgaro JG, Zhu Z. Characterization of the 5-hydroxymethylcytosine-specific DNA restriction endonucleases. Nucleic Acids Res. 2013 Apr;41(7):4198-206. doi: 10.1093/nar/gkt102. Epub 2013 Mar 12.

Bourniquel AA, Bickle TA. Complex restriction enzymes: NTP-driven molecular motors. Biochimie. 2002 Nov;84(11):1047-59.

Burman RW, Yates PA, Green LD, Jacky PB, Turker MS, Popovich BW. Hypomethylation of an expanded FMR1 allele is not associated with a global DNA methylation defect. Am J Hum Genet. 1999 Nov;65(5):1375-86.

Chen IA, Markowitz VM, Chu K, Palaniappan K, Szeto E, Pillay M, Ratner A, Huang J, Andersen E, Huntemann M, Varghese N, Hadjithomas M, Tennessen K, Nielsen T, Ivanova NN, Kyrpides NC. IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res. 2017 Jan 4;45(D1):D507-D516.

Chotai KA, Payne SJ. A rapid, PCR based test for differential molecular diagnosis of Prader-Willi and Angelman syndromes. J Med Genet. 1998 Jun;35(6):472-5. Erratum in: J Med Genet 2000 May;37(5):399.

Dila D, Sutherland E, Moran L, Slatko B, and Raleigh EA. (1990) Genetic and sequence organization of the mcrBC locus of Escherichia coli K-12. J Bacteriol., 172(9):4888-900.

Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M, Sorek R, Rechavi G. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature. 2012 Apr 29;485(7397):201-6.

Dryden DT, Murray NE, Rao DN. Nucleoside triphosphate-dependent restriction enzymes. Nucleic Acids Res. 2001 Sep 15;29(18):3728-41.

Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D Biol Crystallogr. 2010 Apr;66(Pt 4):486-501.

Fouse SD, Nagarajan RO, Costello JF. Genome-scale DNA methylation analysis. Epigenomics. 2010 Feb;2(1):105-17.

Fustin JM, Doi M, Yamaguchi Y, Hida H, Nishimura S, Yoshida M, Isagawa T, Morioka MS, Kakeya H, Manabe I, Okamura H. RNA-methylation-dependent RNA processing controls the speed of the circadian clock. Cell. 2013 Nov 7;155(4):793-806.

Gast FU, Brinkmann T, Pieper U, Krüger T, Noyer-Weidner M, Pingoud A. The recognition of methylated DNA by the GTP-dependent restriction endonuclease McrBC resides in the N-terminal domain of McrB. Biol Chem. 1997 Sep;378(9):975-82.

Hendrickson WA. Anomalous diffraction in crystallographic phase evaluation. Q Rev Biophys. 2014 Feb;47(1):49-93.

Holm L, Rosenström P. Dali server: conservation mapping in 3D. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W545-9.

Hosford CJ, Chappie JS. The crystal structure of the Helicobacter pylori LlaJI.R1 N-terminal domain provides a model for site-specific DNA binding. J Biol Chem. 2018 Jul 27;293(30):11758-11771.

Ishikawa K, Fukuda E, Kobayashi I. Conflicts targeting epigenetic systems and their resolution by cell death: novel concepts for methyl-specific and other restriction systems. DNA Res. 2010 Dec;17(6):325-42.

Jolivet E, L'Haridon S, Corre E, Forterre P, Prieur D. Thermococcus gammatolerans sp. nov., a hyperthermophilic archaeon from a deep-sea hydrothermal vent that resists ionizing radiation. Int J Syst Evol Microbiol. 2003 May;53(Pt 3):847-51.

Jurrus E, Engel D, Star K, Monson K, Brandi J, Felberg LE, Brookes DH, Wilson L,

Chen J, Liles K, Chun M, Li P, Gohara DW, Dolinsky T, Konecny R, Koes DR, Nielsen JE, Head-Gordon T, Geng W, Krasny R, Wei GW, Holst MJ, McCammon JA, Baker NA. Improvements to the APBS biomolecular solvation software suite. Protein Sci. 2018 Jan;27(1):112-128.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015 Jun;10(6):845-58.

Krüger T, Wild C, Noyer-Weidner M. McrB: a prokaryotic protein specifically recognizing DNA containing modified cytosine residues. EMBO J. 1995 Jun 1;14(11):2661-9.

Labrie SJ, Samson JE, and Moineau S. (2010) Bacteriophage resistance mechanisms. Nat Rev Microbiol., 8(5):317-27.

Li F, Zhao D, Wu J, and Shi Y. (2014) Structure of the YTH domain of human YTHDF2 in complex with an m(6)A mononucleotide reveals an aromatic cage for m(6)A recognition. Cell Res., 24(12):1490-2.

Liu K, Ding Y, Ye W, Liu Y, Yang J, Liu J, Qi C. Structural and Functional Characterization of the Proteins Responsible for N(6)-Methyladenosine Modification and Recognition. Curr Protein Pept Sci. 2016;17(4):306-18.

Loenen WA, Dryden DT, Raleigh EA, Wilson GG. Type I restriction enzymes and their relatives. Nucleic Acids Res. 2014 Jan;42(1):20-44.

Luo S, Tong L. Molecular basis for the recognition of methylated adenines in RNA by the eukaryotic YTH domain. Proc Natl Acad Sci U S A. 2014 Sep 23;111(38):13834-9.

Luo GZ, He C. DNA N(6)-methyladenine in metazoans: functional epigenetic mark or bystander? Nat Struct Mol Biol. 2017 Jun 6;24(6):503-506.

Luria SE, Human ML. A nonhereditary, host-induced variation of bacterial viruses. J Bacteriol. 1952 Oct;64(4):557-69.

McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. Phaser crystallographic software. J Appl Crystallogr. 2007 Aug 1;40(Pt 4):658-674. Epub 2007 Jul 13.

Panne D, Raleigh EA, Bickle TA. The McrBC endonuclease translocates DNA in a reaction dependent on GTP hydrolysis. J Mol Biol. 1999 Jul 2;290(1):49-60.

Panne D, Müller SA, Wirtz S, Engel A, Bickle TA. The McrBC restriction endonuclease assembles into a ring structure in the presence of G nucleotides. EMBO J. 2001 Jun 15;20(12):3210-7.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004 Oct;25(13):1605-12.

Pieper U, Schweitzer T, Groll DH, Gast FU, Pingoud A. The GTP-binding domain of McrB: more than just a variation on a common theme? J Mol Biol. 1999b Sep 24;292(3):547-56.

Pieper U, Pingoud A. A mutational analysis of the PD...D/EXK motif suggests that McrC harbors the catalytic center for DNA cleavage by the GTP-dependent restriction enzyme McrBC from Escherichia coli. Biochemistry. 2002 Apr 23;41(16):5236-44.

Santoso B, Ortiz BD, Winoto A. Control of organ-specific demethylation by an element of the T-cell receptor-alpha locus control region. J Biol Chem. 2000 Jan 21;275(3):1952-8.

Sheldrick GM. A short history of SHELX. Acta Crystallogr A. 2008 Jan;64(Pt 1):112-22. Epub 2007 Dec 21.

Stewart FJ, Panne D, Bickle TA, Raleigh EA. Methyl-specific DNA binding by McrBC, a modification-dependent restriction enzyme. J Mol Biol. 2000 May 12;298(4):611-22.

Sukackaite R, Grazulis S, Tamulaitis G, Siksnys V. The recognition domain of the methyl-specific endonuclease McrBC flips out 5-methylcytosine. Nucleic Acids Res. 2012 Aug;40(15):7552-62.

Sutherland E, Coe L, Raleigh EA. McrBC: a multisubunit GTP-dependent restriction endonuclease. J Mol Biol. 1992 May 20;225(2):327-48.

Theler D, Dominguez C, Blatter M, Boudet J, Allain FH. Solution structure of the YTH domain in complex with N6-methyladenosine RNA: a reader of methylated RNA. Nucleic Acids Res. 2014 Dec 16;42(22):13911-9.

Van Duyne GD, Standaert RF, Karplus PA, Schreiber SL, Clardy J. Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin. J Mol Biol. 1993 Jan 5;229(1):105-24.

Wang C, Zhu Y, Bao H, Jiang Y, Xu C, Wu J, Shi Y. A novel RNA-binding mode of the YTH domain reveals the mechanism for recognition of determinant of selective removal by Mmi1. Nucleic Acids Res. 2016 Jan 29;44(2):969-82.

Weigele P, Raleigh EA. Biosynthesis and Function of Modified Bases in Bacteria and Their Viruses. Chem Rev. 2016 Oct 26;116(20):12655-12687.

Xiao CL, Zhu S, He M, Chen, Zhang Q, Chen Y, Yu G, Liu J, Xie SQ, Luo F, Liang Z, Wang DP, Bo XC, Gu XF, Wang K, Yan GR. N(6)-Methyladenine DNA Modification in the Human Genome. Mol Cell. 2018 Jul 19;71(2):306-318.e7.

Xu C, Wang X, Liu K, Roundtree IA, Tempel W, Li Y, Lu Z, He C, and Min J. (2014) Structural basis for selective binding of m6A RNA by the YTHDC1 YTH domain. Nat Chem Biol., 10(11):927-9.

Xu C, Wang X, Liu K, Roundtree IA, Tempel W, Li Y, Lu Z, He C, Min J. Corrigendum: Structural basis for selective binding of m(6)A RNA by the YTHDC1 YTH domain. Nat Chem Biol. 2015 Oct;11(10):815.

Zagorskaitė E, Manakova E, and Sasnauskas G. (2018) Recognition of modified cytosine variants by the DNA-binding domain of methyl-directed endonuclease McrBC. FEBS Lett., 592(19):3335-3345.

Zhang Z, Theler D, Kaminska KH, Hiller M, de la Grange P, Pudimat R, Rafalska I, Heinrich B, Bujnicki JM, Allain FH, Stamm S. The YTH domain is a novel RNA binding domain. J Biol Chem. 2010 May 7;285(19):14701-10.

Zhu T, Roundtree IA, Wang P, Wang X, Wang L, Sun C, Tian Y, Li J, He C, Xu Y. Crystal structure of the YTH domain of YTHDF2 reveals mechanism for recognition of N6-methyladenosine. Cell Res. 2014 Dec;24(12):1493-6.

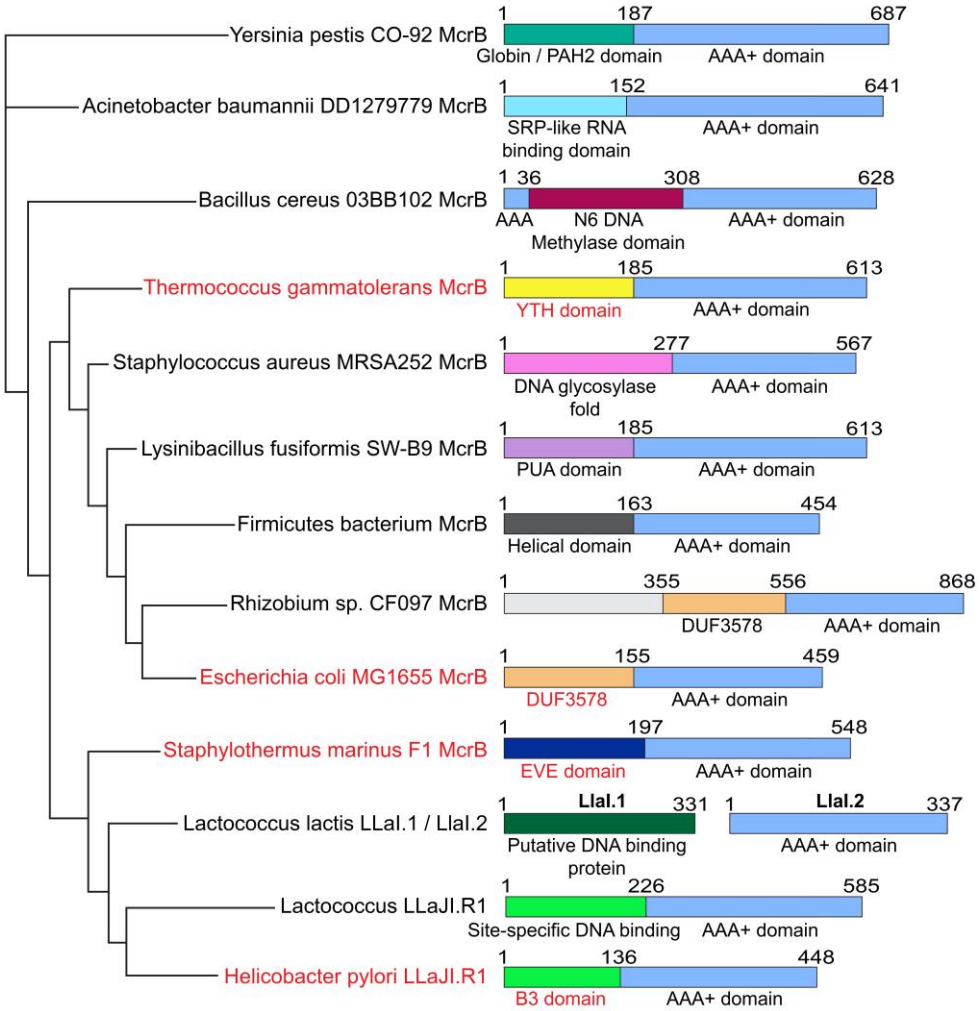

Figure 1. N-terminal domains of McrB homologs are not conserved. Phylogenetic analysis of representative McrB homologs. Conserved C-terminal, GTP-specific AAA+ domains are colored in light blue. Divergent N-terminal domains are colored differently according to the predicted fold. The protein folds of homologs highlighted in red have been experimentally validated by X-ray crystallography. DOE/IMG IDs and applicable PDB codes are as follows: *Yersinia pestis sv. Orientalis* CO-92 McrB, 637199492; *Acinetobacter baumannii* D1279779 McrB, 2563734192; *Bacillus cereus* 03BB102 McrB, 643761466; *Thermococcus gammatolerans* EJ3 McrB, 644807740; *Staphylococcus aureus* MRSA252 McrB, 637153557; *Lysinibacillus fusiformis* SW-B9 McrB, 2598933124; *Firmicutes bacterium* JGI 0000119-P10 McrB, 2519130374; *Rhizobium sp*. CF097 McrB, 2585392831; *Escherichia coli* K-12 MG1655 McrB, 646316336, PDB: 3SSC; *Staphylothermus marinus* F1, DSM 3639 McrB, 640109242, PDB: 6N0S; *Lactococcus lactis lactis* 1AA59 LlaI.1, 263206860; *Lactococcus lactis lactis* 1AA59 LlaI.2, 2632068606; *Lactococcus lactis* LlaJI.R1, 642916737; *Helicobacter pylori* LlaJI.R1, 637022177, PDB: 6C5D.

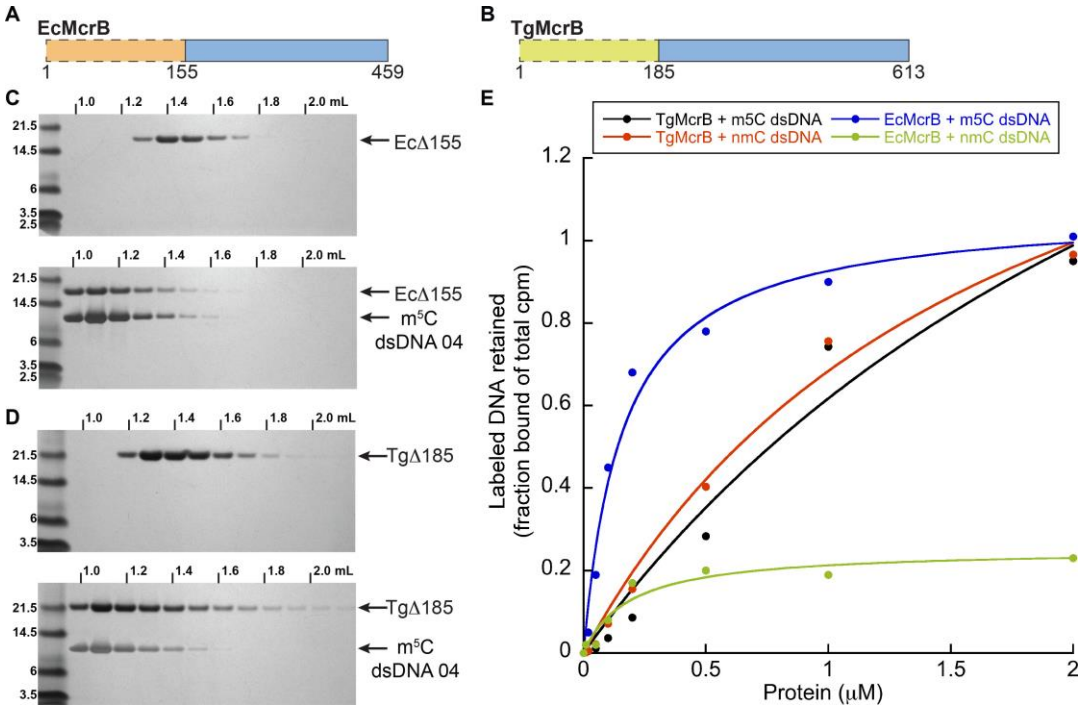

Figure 2. Tgδ185 binds m^{5}C dsDNA. A,B. Domain architectures of EcMcrB (A) and TgMcrB (B) N-terminal domains. Ec DNA binding domain is colored orange and Tg N-terminal domain is colored yellow. The conserved C-terminal AAA+ domain is colored light blue. Truncated constructs used for crystallization and SEC experiments are indicated by the dashed boxes. C. Size shift of Ecδ155 (upper panel) and Ecδ155 + m^{5}C dsDNA (lower panel) are visualized on SDS-PAGE by change in retention volume on SEC. D. Size shift of Tgδ185 (upper panel) and Tgδ155 + m^{5}C dsDNA (lower panel) are visualized on SDS-PAGE by change in retention volume on SEC. Ecδ155 and Tgδ185 are both capable of binding the same m^{5}C DNA substrates as indicated by the respective protein bands size shift to an earlier retention volume. E. Filter binding analysis of TgMcrB and EcMcrB binding to 5-methylctosine modified (m^{5}C) and non-methylated (nm) dsDNA substrates.

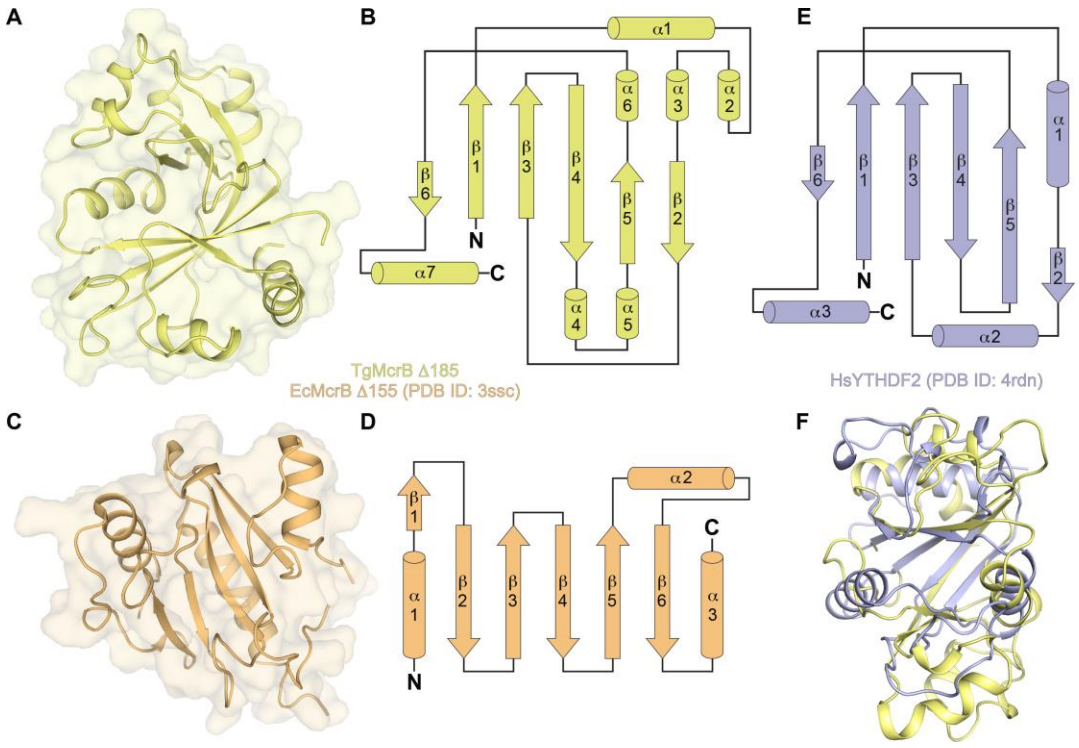

Figure 3. Tgδ185 adopts a YTH-fold and varies from the putative Ecδ155 fold. A, B. Structure (A) and topology (B) of Tgδ185 (yellow). C, D. Structure (C) and topology (D) of Ecδ155 (orange). E. Topology diagram of HsYTHDF2 YTH domain (light blue). F. Structural superposition of Tgδ185 (yellow) and the HsYTHDF2 YTH domain (light blue).

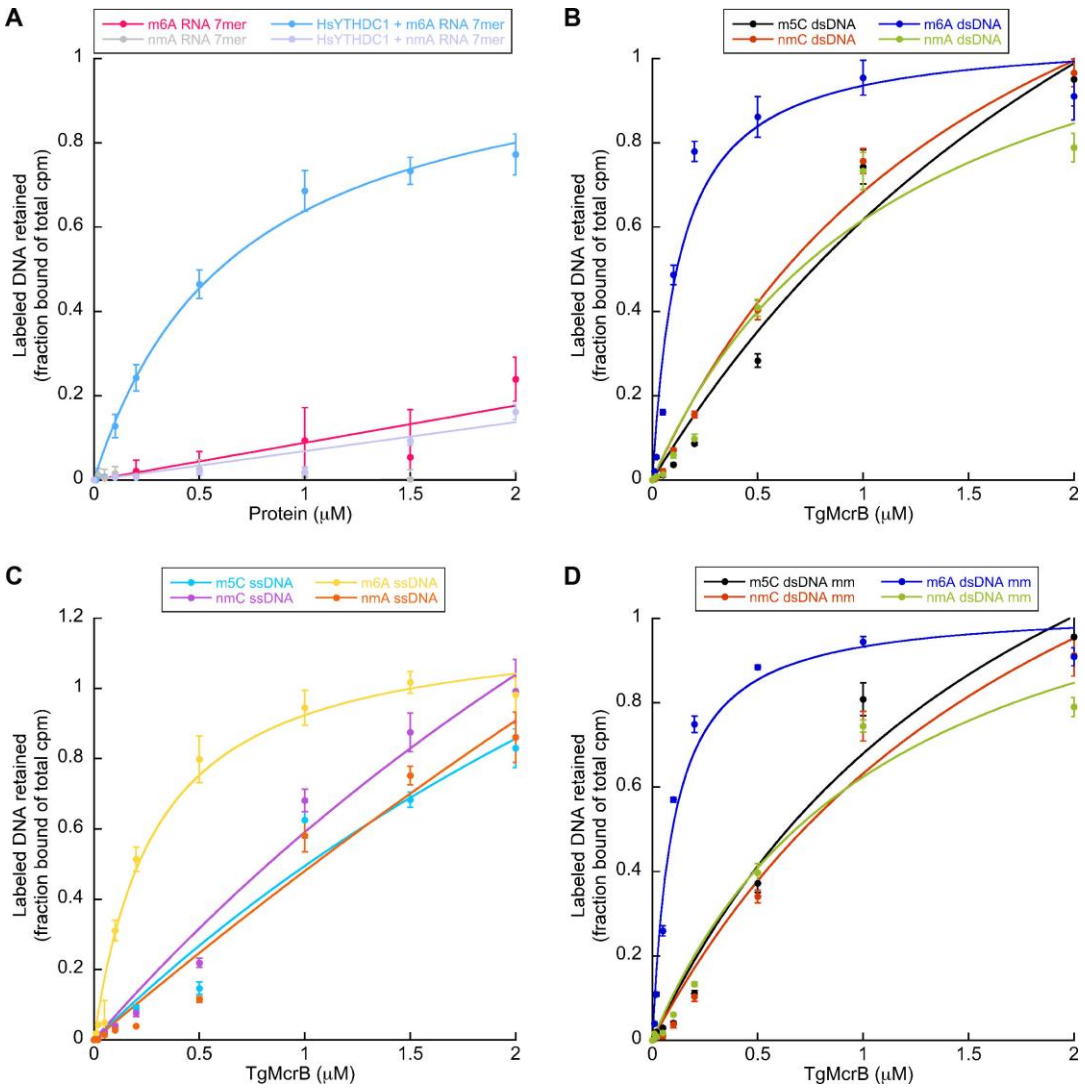

Figure 4. TgMcrB preferentially binds DNA containing m^{6}A modifications. All data represent the average of at least three independent experiments. Calculated K_d values are listed in Supplementary Table S3. m^{5}C and m^{6}A denote 5-methylcytosine and 6-methyladenine modifications respectively. nmC and nmA denote non-methylated versions of the same substrates. A. Filter binding analysis of TgMcrB and HsYTHDC1 YTH domain interactions with RNA substrates. B. Filter binding analysis of TgMcrB interactions with double stranded (ds) DNA substrates. Binding curves from Figure 2E are included for comparison. C. Filter binding analysis of TgMcrB interaction with different single stranded (ss) DNA substrates. D. Filter binding analysis of TgMcrB with different mismatched dsDNA substrates.

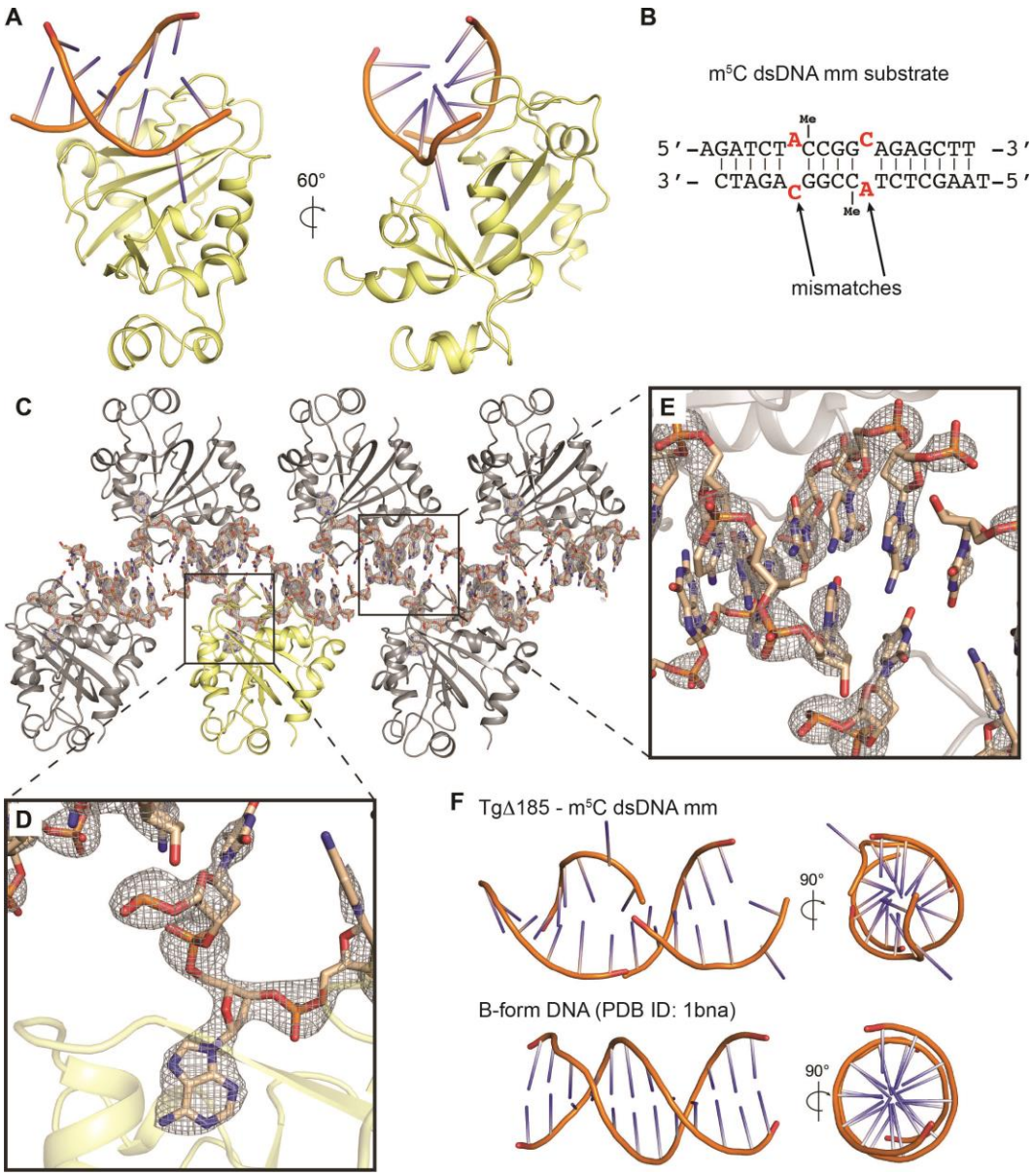

Figure 5. Structure of Tgδ185 bound to m^{5}C dsDNA mm. *A.* Cartoon representation of Tgδ185 bound to m^{5}C dsDNA mm shown in two orientations. Tgδ185 is colored *yellow* and bound DNA is colored *wheat*. *B.* Schematic of the m^{5}C dsDNA mm substrate used for crystallization with Tgδ185. Mismatched bases are colored red and indicated by arrows. *C.* Crystal packing of Tgδ185 with m^{5}C dsDNA mm. One asymmetric unit is colored *yellow* with the bound DNA illustrated as sticks and colored *wheat*. The electron density map of the DNA is colored light grey and illustrated as mesh. *D.* Zoomed in view of the electron density surrounding the flipped-out adenine base. *E.* Zoomed in view of the electron density surrounding base pairs within the bound DNA duplex. *F.* Structural comparison of m^{5}C dsDNA mm to B-form DNA (PDB ID: 1bna) illustrates deformation in the bound DNA.

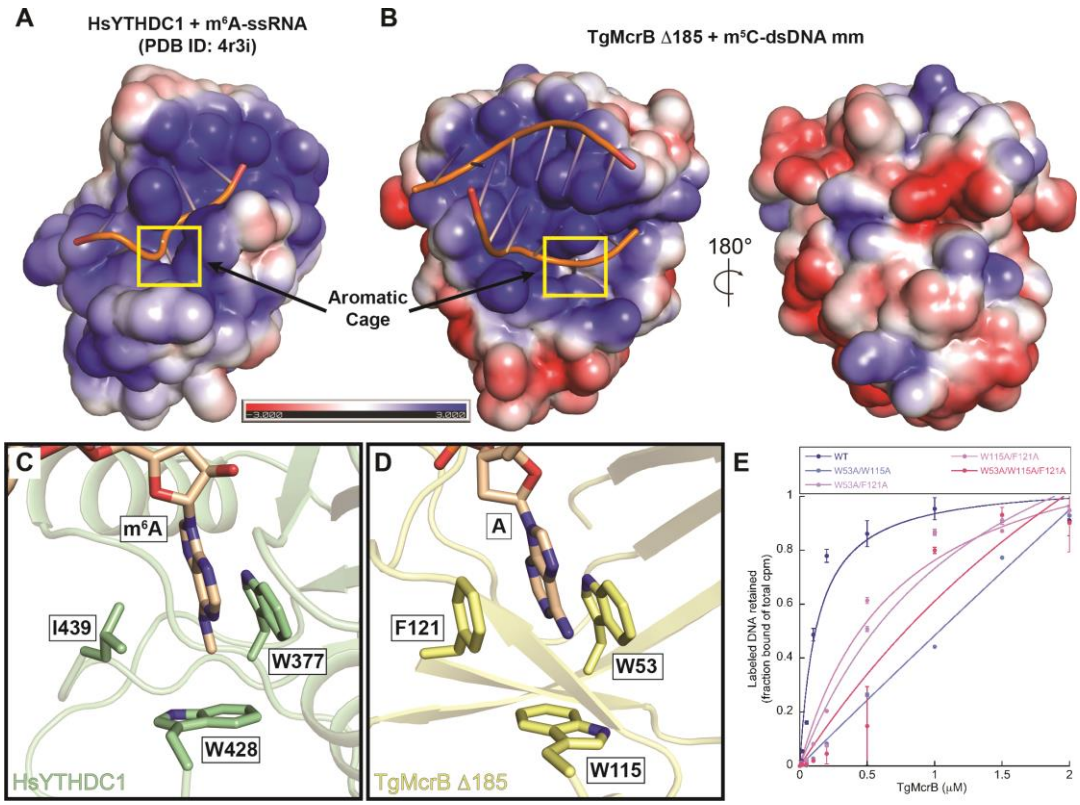

Figure 6. Tgδ185 utilizes a structurally conserved aromatic cage to bind DNA. *A.* Electrostatic surface of HsYTHDC1 + m^{6}A-ssRNA. *B.* Electrostatic surface of Tgδ185 + m^{6}A-dsDNA mm. A *yellow* box is drawn around the aromatic cage and indicated by arrows. *Scale bar* indicates electrostatic surface coloring from −3 K_bT/e_c to +3 K_bT/e_c. *C.* Zoomed in view of the HsYTHDC1 aromatic cage residues (*green*) and m^{6}A base (*wheat*). *D.* Zoomed in view of the Tgδ185 aromatic cage residues (*yellow*) and modelled adenine base (*wheat*). *E.* Filter binding analysis of TgMcrB wild-type and aromatic cage mutants with m^{6}A dsDNA.

Supplementary information for:

The crystal structure of the *Thermococcus gammatolerans* McrB N-terminal domain defines a new mode of substrate recognition and specificity among McrB homologs

Christopher J. Hosford[1] and Joshua S. Chappie[1,*]

[1] Department of Molecular Medicine, Cornell University, Ithaca, NY, 14853, USA

* To whom correspondence should be addressed. Tel: +1 (607) 253-3654; Fax: +1 (607) 253-3659; Email: chappie@cornell.edu

Supplementary Tables

Supplementary Figures

Table S1. DNA and RNA oligonucleotide sequences.

m5C 04 us	5’-CCGGGTAAGA(m^5C)CGGTAGCGAGCCCCGAGCGAT (m^5C)CGGAGAATGGGCC-3’
m5C 04 ls	5’-GGCCCATTCT(m^5C)CGGATCGCTCGGGGCTCGCTA (m^5C)CGGTCTTACCCGG-3’
m5C us	5'-AGATCTA(m^5C)CGGTAGAGCTT-3'
m5C ls	5'-TAAGCTCTA(m^5C)CGGTAGATC-3'
nmC us	5'-AGATCTACCGGTAGAGCTT-3'
nmC ls	5'-TAAGCTCTACCGGTAGATC-3'
m6A us	5'-AGATCTA(m^6A)CGGTAGAGCTT-3'
m6A ls	5'-TAAGCTCTA(m^6A)CGGTAGATC-3'
nmA us	5'-AGATCTAACGGTAGAGCTT-3'
nmA ls	5'-TAAGCTCTAACGGTAGATC-3'
m5C mm us	5’-AGATCTA(m^5C)CGGCAGAGCTT-3’
m5C mm ls	5'-TAAGCTCTA(m^5C)CGGCAGATC-3'
nmC mm us	5’-AGATCTACCGGCAGAGCTT-3’
nmC mm ls	5'-TAAGCTCTACCGGCAGATC-3'
m6A mm us	5'-AGATCTA(m^6A)CGTCAGAGCTT-3'
m6A mm ls	5'-TAAGCTCTA(m^6A)CGTCAGATC-3'
nmA mm us	5'-AGATCTAACGTCAGAGCTT-3'
nmA mm ls	5'-TAAGCTCTAACGTCAGATC-3'
m6A RNA 7mer	5’-CGG(m^6A)CUG-3’
nmA RNA 7mer	5’-CGGACUG-3’

Table S2. X-ray Data collection and refinement statistics.

Data collection			
Model	**TgΔ185 Apo**	**TgΔ185 + meDNA 1**	**TgΔ185 + meDNA 2**
PDB code	XXXX		XXXX
X-ray Source	NECAT 24ID-C	NECAT 24ID-C	NECAT 24ID-E
Wavelength (Å)	0.9791	0.9791	0.9791
Spacegroup	C2	$P2_12_12_1$	$P2_12_12_1$
a, b, c (Å)	67.84, 43.99,	41.87, 56.50,	41.17, 57.30,
α, β, γ (°)	90.00, 120.28,	90.00, 90.00,	90.00, 90.00,
Resolution (Å)[a]	53.51 – 1.68	109.28 – 2.64	107.51 – 2.27
No. measured	214,344 (2,293)	76,849 (7,731)	430,541 (64,292)
No. unique	17,622 (646)	7,972 (895)	17,710 (2,477)
Completeness (%)[a]	97.7 (70.7)	98.0 (85.3)	99.4 (100.0)
Multiplicity[a]	12.2 (3.5)	9.6 (8.6)	24.3 (26.0)
R_{meas}[a]	0.069 (0.127)	0.066 (3.016)	0.082 (1.349)
Mean I/σ_I[a]	35.0 (7.2)	20.8 (0.6)	18.0 (2.59)
$CC_{1/2}$[a]	0.999 (0.987)	0.999 (0.400)	0.999 (0.942)
Phasing			
Initial F.O.M			
No. Se sites			
Refinement			
R_{work}/R_{free}	0.1770/0.1907		0.2455/0.2851
RMSD			
Bond lengths	0.013		0.010
Bond angles (°)	1.46		1.28
Ramachandran plot			
Favored (%)	98.27		97.66
Allowed (%)	1.73		2.34
Outliers (%)	0.00		0.00
Average B-Factor	30.42		79.09
Clashscore	4.51		8.27
No. Atoms			
Macromolecule			
DNA			
Solvent			

[a] Denotes values for the highest resolution shell

Table S3. Dissociation constants from filter binding experiments

Construct	DNA or RNA	Kd (μM)	Error (%)
EcMcrB WT	m^5C dsDNA	0.1395	0.4747
EcMcrB WT	nmC dsDNA	ND*	ND*
TgMcrB WT	m^5C dsDNA	0.7619	3.0657
TgMcrB WT	nmC dsDNA	0.6299	2.9526
TgMcrB WT	m^6A dsDNA	0.1165	2.4366
TgMcrB WT	nmA dsDNA	0.6987	1.0414
TgMcrB WT	m^6A RNA 7mer	ND*	ND*
TgMcrB WT	nm RNA 7mer	ND*	ND*
HsYTHDC1	m^6A RNA 7mer	0.5931	2.6551
HsYTHDC1	nm RNA 7mer	ND*	ND*
TgMcrB WT	m^5C dsDNA mm	0.6426	3.3955
TgMcrB WT	nmC dsDNA mm	0.7176	2.6344
TgMcrB WT	m^6A dsDNA mm	0.0953	2.4667
TgMcrB WT	nmA dsDNA mm	0.6791	3.7252
TgMcrB WT	m^5C ssDNA (US)	1.0117	1.1252
TgMcrB WT	nmC ssDNA (US)	0.8253	0.1404
TgMcrB WT	m^6A ssDNA (US)	0.2108	0.6875
TgMcrB WT	nmA ssDNA (US)	1.0424	5.9554
TgMcrB W53A/W115A	m^6A dsDNA	1.0341	1.3319
TgMcrB W53A/F121A	m^6A dsDNA	0.5496	4.0478
TgMcrB W115A/F121A	m^6A dsDNA	0.4449	3.9907
TgMcrB W53A/W115A/F121A	m^6A dsDNA	0.7823	5.5741
TgMcrB E17A/N19A	m^6A dsDNA	0.0773	1.0464
TgMcrB Y61A/N82A	m^6A dsDNA	0.0175	10.939
TgMcrB R78A/R81A	m^6A dsDNA	0.0952	0.3428

*ND – signifies not determined due to incomplete saturation within the data acquisition range.

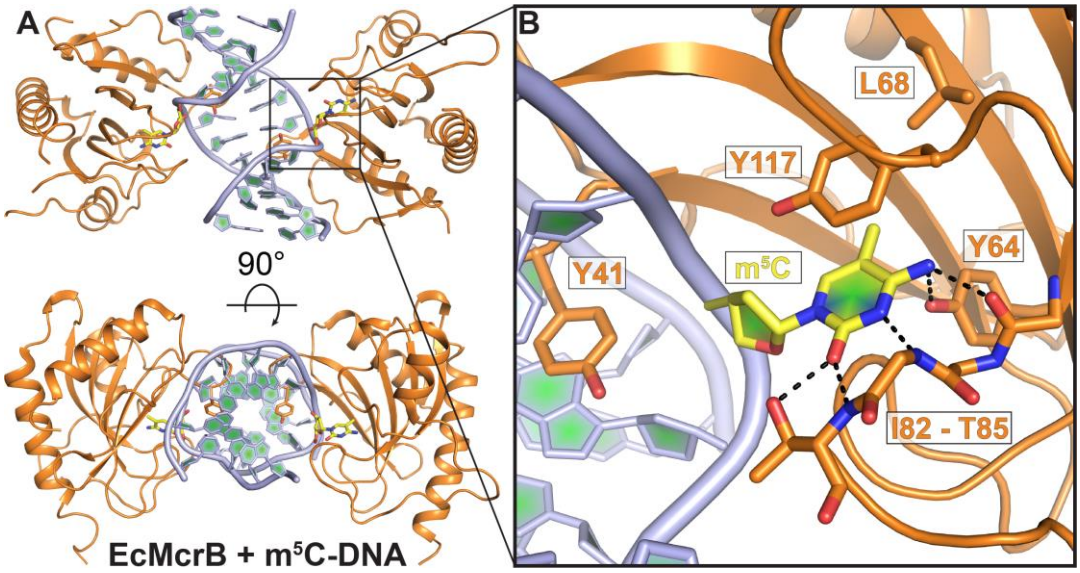

Figure S1. EcMcrB recognizes methylated DNA via base flipping. A. Cartoon representation of Ecδ155 (orange) bound to m^5C-DNA (light blue) shown in two, perpendicular orientations (PDB: 3SSC). B. Zoomed in view of the binding pocket in Ecδ155 that recognizes the flipped out m^5C base (yellow). Residues contacting the base are labeled along with the inserted tyrosine (Y41) that stabilizes the gap in the duplex through base stacking. Dashed black lines indicate hydrogen bonds.

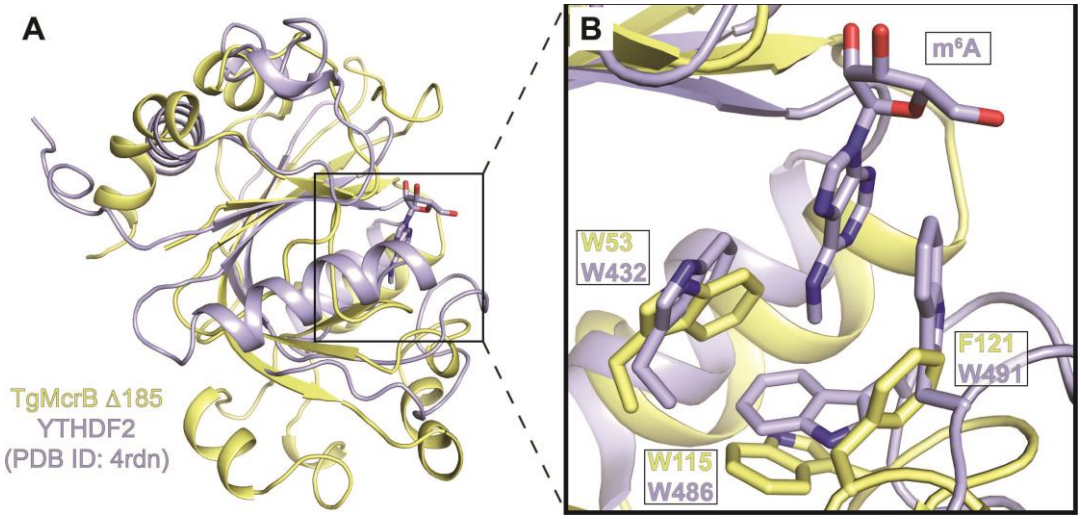

Figure S2. Structural superposition of Tgδ185 with HsYTHDF2. A. Structural superposition of Tgδ185 (yellow) with HsYTHDF2 (light blue) shown in cartoon representation. The m^6A base bound to HsYTHDF2 is shown in sticks (light blue). B. Zoomed in view of the HsYTHDF2 aromatic cage surrounding the bound m^6A base (W432, W486, and W491). The aromatic cage residues are structurally conserved in Tgδ185 (W53, W115, F121).

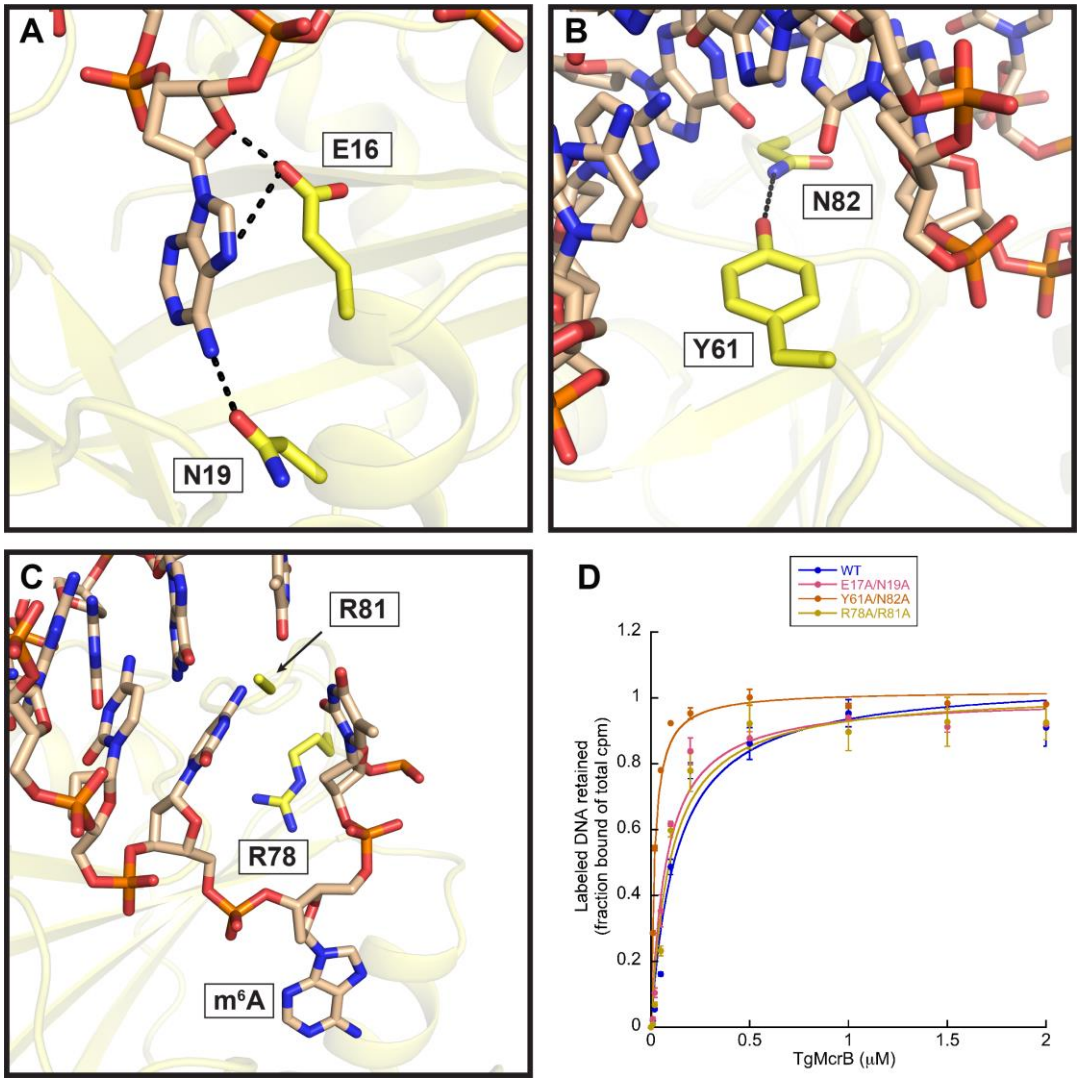

Figure S3. Additional Tgδ185-DNA interactions. A. Residues E16 and N19 form electrostatic interactions with the modelled adenine base that is flipped out of the DNA duplex into the aromatic cage. B. Residues Y61 and N82 form a 'wedge' within the major groove. C. R78 and R81 (unmodelled) are possible candidates for forming base-specific contacts to DNA. D. Filter binding analysis of E16A/N19A, Y61A/N82A, and R78A/R81A double mutants. Binding was carried out using m^6A dsDNA substrates. All data represent the average of at least three independent experiments. Calculated K_d values are listed in Supplementary Table S3.

Chapter 4. The N-terminal domain of *Staphylothermus marinus* McrB shares structural homology with PUA-like RNA binding proteins

The N-terminal domain of *Staphylothermus marinus* McrB shares structural homology with PUA-like RNA binding proteins

Christopher J. Hosford[1], Yiming Niu[1], and Joshua S. Chappie[1,*]

From the [1]Department of Molecular Medicine, Cornell University, Ithaca NY 14853

*To whom correspondence should be addressed: Joshua S. Chappie: Department of Molecular Medicine, Cornell University, Ithaca NY 14853; chappie@cornell.edu; Tel. (607) 253-3654; Fax. (607) 253-3659.

ABSTRACT

McrBC is a conserved modification-dependent restriction system that in *Escherichia coli* specifically targets foreign DNA containing methylated cytosines. Recent crystallographic data show that the N-terminal domain of *Escherichia coli* McrB binds substrates via a base flipping mechanism. This region is poorly conserved among the plethora of McrB homologs, suggesting that other species may use alternative binding strategies and/or recognize different targets. Here we present the crystal structure of the N-terminal domain from *Stayphlothermus marinus* McrB (Sm3-180) at 2.10Å, which adopts a PUA-like EVE fold that is closely related to the YTH and ASCH RNA binding domains. Unlike most PUA-like domains, Sm3-180 preferentially binds DNA and can associate with different modified substrates. Structural comparison shows that the canonical 'aromatic cage' binding pocket present in other EVE/YTH domains is degenerate in Sm3-180, which may explain its promiscuity in target recognition. Mutagenesis and filter binding support this hypothesis. We also identify a specific helical insert present in subset of PUA-like domains that correlates with the ability to bind DNA. Together these data have important implications for PUA-like domain specificity and suggest a broader biological versatility for the McrBC family than

previously described.

INTRODUCTION

Restriction modification systems (RMS) are conserved defense systems that protect bacteria against viral bacteriophage (phage) infection (Labrie et al., 2010). Classical RMS contain a site-specific DNA binding module, endonuclease core, and associated methyltransferase to protect the host genome (Tock and Dryden, 2005). As phages incorporated modifications into their genomes to evade RMS, bacteria in turn evolved modification-dependent restriction systems (MDRS) – that specifically target and cleave methylated and/or glucosylated DNA – to restore the balance in the ongoing arms race for survival (Loenen and Raleigh, 2014). These systems define the epigenetic landscape of bacterial populations (Ishikawa et al., 2010) and complement CRISPR-Cas systems as essential barriers to foreign invaders (Dupuis et al., 2013).

McrBC is a highly conserved, two-component MDRS consisting of the McrB and McrC proteins. *Escherichia coli* (*Ec*) McrB contains an N-terminal DNA binding domain that recognizes 4-methyl-, 5-methyl-, or 5-hydroxymethylcytosines (Sutherland et al., 1992; Krüger et al., 1995; Gast et al., 1997) and a C-terminal AAA+ domain that hydrolyzes GTP and facilitates nucleotide-dependent oligomerization (Panne et al., 2001). EcMcrC contains a C-terminal PD-(D/E)XK nuclease domain but cannot bind DNA on its own (Pieper and Pingoud, 2002). To exert its function, EcMcrC associates with the EcMcrB oligomer (Pieper and Pingoud), which *in vitro* stimulates the GTP hydrolysis and subsequent DNA translocation (Pieper et al., 1999; Panne et al., 1999). Collision of two McrBC complexes is thought to trigger DNA cleavage on both strands

near the modified sites (Steward et al., 2000; Pieper et al., 2002). These mechanochemical properties are reminiscent of type I and type III RMS, which bind DNA at non-modified sites separated by up to thousands of base pairs and use ATP hydrolysis to power similar long-range translocation events through which cleavage occurs either by collision or stalling (Dryden et al., 2001). Despite this similarity, the structural and molecular details of McrBC assembly and translocation remain poorly defined and it remains to be seen whether these mechanistic features are conserved in other organisms beyond *E. coli*.

Recent crystallographic data shows that EcMcrB binds modified DNA via a base-flipping mechanism (Sukackaite et al., 2012). The N-terminal domain of EcMcrB, however, is poorly conserved among the wide array of McrBC homologs, suggesting that other species may use different mechanisms for substrate binding and/or may preferentially target other sequences and modifications. This remains a largely unexplored area of study. Here we present the crystal structure of the N-terminal domain from *Stayphlothermus marinus* McrB (Sm3-180) at 2.10 Å, which adopts a **P**seudo**U**ridine synthase and **A**rchaeosine transglycosylase (PUA)-like EVE domain fold that is prevalent among prokaryotic RNA binding proteins and shares homology with eukaryotic YTH/ASCH family proteins. Sm3-180, however, preferentially binds DNA and associates with different modified substrates. Structural comparison shows that the canonical 'aromatic cage' binding pocket found PUA-like domains is degenerate in Sm3-180, which may explain its promiscuity in target recognition. Mutagenesis and filter binding support this hypothesis. We also identify a specific helical insert present in subset of PUA-like domains that correlates with the ability to

bind DNA. Together these data have important implications for PUA-like domain specificity and suggest new mechanistic possibilities for McrBC enzymes, underscoring the modular nature of these nuclease complexes.

RESULTS

SmMcrB 3-180 adopts an EVE domain fold

Previous biochemical and structural studies established that the EcMcrB N-terminal domain (residues 1-155) recognizes DNA containing methylated cytosines (5mC) (Sutherland et al., 1992; Krüger et al., 1995; Gast et al., 1997) via a base flipping mechanism (Sukackaite et al., 2012). This domain, however, is only conserved in a handful of McrB homologs (Sukackaite et al., 2012), suggesting other species use different strategies for substrate binding and/or may preferentially target other sequences and modifications. To test this hypothesis and explore the evolutionary diversity of this family, we screened divergent McrB homologs containing unique N-terminal sequences and identified the N-terminal domain from *Staphylothermus marinus* McrB (Sm3-180; Figure 1A) as a suitable candidate for structural and biochemical characterization. This construct is thermally stable, could be expressed in *E. coli* and purified to homogeneity in milligram quantities, and readily crystallized by sitting drop vapor diffusion. The C-terminal AAA+ domain of SmMcrB and the accompanying SmMcrC nuclease share identifiable homology with their *E. coli* counter parts (Figure 1A), suggesting that only the putative substrate binding module is distinct in this species while the motor and cleavage machinery remain unaltered. Recombinant selenomethionine-labeled Sm3-180 yielded crystals of the space group $P4_32_12$ with 1

molecule in the asymmetric unit. The structure was solved by single wavelength anomalous diffraction (SAD) phasing (Hendrickson, 2014) and the final model was refined to 2.10Å resolution with R_{work} and R_{free} values of 0.1880 and 0.2298 (Table 1).

Sm3-180 is composed of a central six-stranded pseudobarrel, connected with an intricate network of extended loops and α-helical inserts (Figure 1B). The strands are ordered β6-1-3-4-5-2 with each oriented in an antiparallel configuration except β1 and β3, which are parallel (Figure 1B). Helix α1 lies between β2-β3 while helices α2 and α3 are inserted in tandem between β4-β5 (Figure 1C). A fourth helix, α4, follows β6 at the C-terminal end of the domain (Figure 1C). Attempts to superimpose Sm3-180 with N-terminal DNA binding domain of EcMcrB (PDB: 3SSD) failed to yield a consistent structural alignment, suggesting their structural topologies differ significantly. A homology search via the Dali server (Holm and Rosenström, 2010) supports this presumption and instead classifies Sm3-180 as an EVE domain fold. EVE domains are part of a larger superfamily of PUA-like domains that also include YTH and ASCH folds (Bertonati et al., 2009). These domains all share a core five-stranded pseudobarrel architecture and function as RNA recognition modules in bacterial, archaeal, and eukaryotic proteins (Iyer et al., 2006; Perez-Arellano, 2007; Bertonati, et al. 2009). Although many PUA-like domains also contain various inserts spaced throughout the core fold, EVE and YTH domains contain a sixth β-strand and are more closely related despite low sequence identity (Bertonati, et al. 2009). Structural superposition of Sm3-180 with three of the top Dali hits confirms structural similarity with PUA-like domain containing proteins: the PSPTO5229 EVE domain (PDB: 2eve, Z-score = 9.8, RMSD = 2.9Å), the *Zymomonas mobilis* (Zm) ASCH domain (PDB: 5y6c, Z-score = 10.9,

RMSD = 2.6Å), and YTHDC1 YTH domain (PDB: 4r3i, Z-score = 9.1, RMSD = 2.6Å) (Figure 2A). A structure-based sequence alignment further shows that Sm3-180 contains all the residues characteristic of prokaryotic EVE domains, distinguishing it from eukaryotic YTH domains (Figure 2B).

SmMcrB preferentially binds DNA

PUA-like domains bind RNA and contain a swath of positively charged residues on one face that forms a cleft for the negatively charged RNA phosphate backbone (Figure 3A). Two sulfate ions co-crystallized with Sm3-180, which are bound in this cleft and localized along the positively charged electrostatic surface (Figure 3A). Of all previously identified PUA-domains, only YTH domains have been extensively studied and specifically associate with short RNAs containing 6-methyladenine (m^6A), with G(m^6A)C serving as the primary consensus site (Xu et al., 2015). EVE domains remain largely uncharacterized and minimal biochemical data exists supporting their RNA binding capacity. To determine whether SmMcrB shares this RNA binding activity, we analyzed Sm3-180's association with methylated RNA oligonucleotides via filter binding (Fig. 3B, Supplementary Table 1). In contrast to YTH domains (Luo and Tang, 2014; Xu, et al., 2014; Theler, et al., 2014; Xu, et al., 2015), Sm3-180 shows only weak affinity for m^6A RNA (Fig. 3B, cyan and purple). Recent structural studies demonstrated that a putative *Zymomonas mobilis* ASCH protein (ZmASCH) can bind DNA as well as RNA (Kim et al., 2017). We therefore tested whether Sm3-180 could interact with methylated (m^6A or m^5C) and/or non-methylated DNA substrates (Supplementary Table 1). Sm3-180 binds DNA with much stronger affinity and does

not discriminate between different modifications (Fig. 3B, black, red, and blue). We also note that Sm3-180 binds single stranded m^6A DNA with nearly the same affinity as the double stranded substrates (Figure 3B). These data argue that Sm3-180 is an EVE domain that preferentially binds DNA.

SmMcrB contains a degenerate aromatic cage and large β4-β5 insert

Given that all other PUA-like domains show a strong preference for RNA, we examined the available structural models in greater detail to identify topological features that could explain Sm3-180's unique substrate binding profile. A key defining feature of all PUA-like domains is the presence of a conserved hydrophobic pocket – colloquially termed the 'aromatic cage' – that sits at the base of the positively charged cleft (Figure 3A, yellow circles). Structural characterization of YTH domain-RNA complexes (Luo and Tang, 2014; Xu et al., 2014; Theler et al., 2014; Xu et al., 2015) has shown that the aromatic cage is critical for substrate binding and discrimination, stabilizing the m^6A base through a combination of hydrophobic interactions and π-π stacking (Figure 4A). In the human YTHDC1 YTH domain, W377 in β2 and W428 and L439 in the β4-β5 loop form the cage (Figure 4A, 5A). This canonical arrangement is also observed in the PSPTO5229 EVE domain structure with F13 from β1-α1 loop, W25 from β2-β3 loop, and Y82 from β4-β5 loop occupying similar spatial positions (Figure 4B, 5A). Sm3-180 and ZmASCH both contain a helical insert (α2-α3) within the β4-β5 loop that is absent in PSPTO5229 and YTHDC1 (Figure 5B). These inserts alter the overall shape and organization of the aromatic cage, causing them to deviate from the canonical arrangement. In ZmASCH, W15 from the β1-α1 loop, F18 from α1, and Y90 from α3

in the β4-β5 insertion form the cage (Figure 4C, 5A). Structural superposition aligns W31 and Y164 in Sm3-180 with W428 and W377 from YTHDC1 (Figure 4D). SmMcrB lacks an analogous side chain at the position of L439 and instead appears to utilize I123 from α3 to close the binding pocket from the opposite side (Figure 4D). A triple mutant of W31A/I123A/W164A retains dsDNA binding capacity albeit at a reduced level (Figure 4E), suggesting the Sm3-180 aromatic cage is degenerate and that other structural motifs may contribute to substrate recognition.

DISCUSSION

Here we showed that the N-terminal domain of SmMcrB is an EVE domain. Although EVE domains are distinct among PUA-like domains, the overall topological similarities to YTH and ASCH domains allow for structural and functional comparisons. YTH domains preferentially bind m^6A modified RNA, with the conserved aromatic cage residues serving as the sole determinants of substrate recognition and discrimination (Luo and Tang, 2014; Xu et al., 2014; Theler et al., 2014; Xu et al., 2015). Our data shows Sm3-180 preferentially binds DNA (Figure 3B). Structural superposition reveals that the Sm3-180 aromatic cage is degenerate and deviates from the canonical arrangement seen in other EVE and YTH domains (Figure 4). Mutation of the aromatic cage in Sm3-180 only partially inhibits DNA binding (Figure 4E), suggesting an alternate mode of substrate recognition is utilized.

Sm3-180 has a large helical insert in the β4-β5 loop (Figure 5), which changes the organization of the aromatic cage and may provide additional motifs that confer specificity for DNA. Consistent with this hypothesis is the observation that ZmASCH

also contains a helical insert (Figure 5B) and can bind both DNA and RNA (Kim et al., 2017). The ZmASCH aromatic cage similarly deviates from the canonical arrangement, with two of the three hydrophobic residues localizing to the opposite side of the central β-sheet (Figure 5A). Moreover, the recently deposited structure of the human THYN1 protein in complex with m^5C DNA reveals an EVE domain with a similar helical insert in the β4-β5 loop (PDB: 5j3e). THYN1 primarily interacts with DNA via various polar and charged residues on the basic patch with the insert reaches around to contact phosphate backbone. The presence of a β4-β5 loop insert thus appears to correlate with the ability to bind DNA and may serve as an important predictive feature when characterizing new PUA domain-containing proteins.

Sm3-180 constitutes the substrate binding domain of *Staphylothermus marinus* McrB. Despite its preference for DNA, Sm3-180 displays promiscuity with regard to binding modifications (Figure 3B). This is in stark contrast to the N-terminal domain of *E. coli* McrB, which strictly recognizes DNA containing methylated cytosines (Sutherland et al., 1992; Krüger et al., 1995; Gast et al., 1997). These differences may reflect distinct evolutionary pressures, such as attack by lytic bacteriophages with different genomic content and modifications (Weigele and Raleigh, 2016). Distantly related McrB homologs like LlaJI, LlaI, and BsuMI target DNA site-specifically – a direct consequence of the unmodified viruses they provide protection against (O'Sullivan et al., 1995; Ohshima et al., 2002; O'Driscoll et al., 2006). We recently showed that N-terminal domain of LlaJI.R1 from *Helicobacter pylori* adopts a B3 domain fold to recognize DNA independent of modifications (Hosford and Chappie, 2018). Collectively these observations suggest an emerging theme in which bacteria

have adapted a conserved set of core machinery – the GTP-specific AAA+ motor of McrB and the associated McrC nuclease – to different biological contexts through the incorporation of alternative N-terminal binding domains. Although EcMcrBC is historically described as a prototypical MDRS (Loenen and Raleigh, 2014), it appears members of the McrBC superfamily can be classified more broadly as modular nucleases and can be tuned for different substrates.

EXPERIMENTAL PROCEDURES

Cloning, expression and purification SmMcrB 3-180 constructs

DNA encoding the *Staphylothermus marinus F1* McrB protein (DOE IMG/M ID 640109242; Chen et al., 2017) was codon optimized for *E. coli* expression and synthesized commercially by Integrated DNA Technologies (IDT), Inc. DNA encoding the N-terminal domain (Sm3-180) was amplified by PCR and cloned into pCAV4, a modified T7 expression vector that introduced an N-terminal 6xHis-NusA tag followed by a Hrv3C protease site upstream of the inserted sequence. Seleno-methionine labeled (SeMet) Sm3-180 was transformed into BL21(DE3) cells, grown at 37°C in minimal media, and expressed in the absence of auxotrophs as described previously (Van Duyne et al., 1993). Native Sm3-180 was transformed into BL21(DE3) cells, grown at 37°C in Terrific Broth to an OD_{600} of 1.0, and then induced with 0.3 mM IPTG overnight at 19°C. All cells were harvested, washed with nickel load buffer (20 mM HEPES pH 7.5, 500 mM NaCl, 30 mM imidazole, 5% glycerol (v:v), and 5 mM β-mercaptoethanol), and pelleted a second time. Pellets were typically flash frozen in liquid nitrogen and stored at -80°C.Thawed pellets from 500 ml cultures were resuspended in 30 ml of

nickel load buffer supplemented with 10 mM PMSF, 5 mg DNAse I (Roche), 1 mM MgCl2, and a complete protease inhibitor cocktail tablet (Roche). Lysozyme was added to 1 mg/ml and the mixture was incubated for 15 minutes rocking at 4°C. Cells were disrupted by sonication and the lysate was cleared of debris by centrifugation at 13 000 rpm (19 685 g) for 30 minutes at 4°C. For native and SeMet Sm3-180, the supernatant was filtered, loaded onto a 5 ml HiTrap chelating column charged with $NiSO_4$ and then washed with nickel load buffer. Sm3-180 was eluted with an imidazole gradient from 30 mM to 1 M. Hrv3C protease was added to pooled fractions and dialyzed overnight at 4°C into SP loading buffer (20 mM HEPES pH 7.5, 50 mM NaCl, 1 mM EDTA, 5% glycerol (v:v), and 5 mM DTT). The sample was applied to a 5 ml HiTrap SP HP column equilibrated with SP loading buffer and then washed with SP loading buffer. Sm3-180 was eluted with a NaCl gradient from 50 mM to 1 M. Pooled fractions were subjected to a 30 kDa Millipore centrifugal concentrator, flow through collected, and concentrated on a 10 kDa centrifugal concentrator. The concentrated protein was further purified by size exclusion chromatography (SEC) using a Superdex 75 10/30 pg column. All proteins were exchanged into a final buffer of 20 mM HEPES pH 7.5, 150 mM KCl, 5 mM MgCl2, and 1 mM DTT (5 mM for SeMet labelled) during SEC and concentrated to 5-40 mg/ml. Concentrations of purified proteins were determined by SDS-PAGE with BSA standards. All point mutations were introduced into Sm3-180 in pCAV4 by quick change PCR and proteins purified as described previously.

Crystallization, X-ray data collection, and structure determination

SeMet Sm3-180 was crystallized by sitting drop vapor diffusion in 0.1 M BisTris

Propane pH 7.5, 0.2 M Na_2SO_4, and 23% PEG3350 with a drop size of 2 μL and reservoir volume of 65 μL. The reservoir was supplemented with 5 mM DTT immediately prior to setting up the drop. Crystals typically appeared within 2-8 days at 20°C and were cryoprotected with Parabar 10312 and frozen in liquid nitrogen. They were of the space group $P4_32_12$ with unit cell dimensions a = 62.41 Å, b = 62.41 Å, c = 118.63 Å and α = 90.00°, β = 90.00°, γ = 90.00°. Single-wavelength anomalous diffraction (SAD) data were collected remotely on the tuneable NE-CAT 24-ID-C beamline at the Advanced Photon Source at the selenium edge energy at 12.663 kEv (Table 1). Data were integrated and scaled using XDS (Kabsch, 2010) and AIMLESS (Evans, 2006) via the NE-CAT RAPD pipeline. Heavy atom sites were located using SHELX (Sheldrick, 2008) and phasing, density modification, and initial model building was carried out using the Autobuild routines of the PHENIX package (Adams et al., 2010). Further model building and refinement was carried out manually in COOT (Emsley et al., 2010) and PHENIX (Adams et al., 2010) respectively. The final model was refined to 2.10Å resolution with Rwork/Rfree = 0.1880/0.2298 (Table 1) and contained one molecule in the asymmetric unit: chain A, 3-180. All structural models were rendered with Pymol (http://www.pymol.org) and surface electrostatics were calculated with APBS (Jurrus et al., 2018).

Preparation of oligonucleotide substrates

DNA (Integrated DNA technologies, IDT) and RNA (Dharmacon) for filter binding were synthesized commercially as lyophilized, single-stranded oligonucleotides. All oligonucleotides were resuspended to 1 mM in 10 mM Tris-HCl and 1 mM EDTA and

stored at -20°C until needed. Single-stranded oligonucleotides were 5' end-labeled with (γ32P)ATP using polynucleotide kinase (New England Biolabs) and then purified on a P-30 spin column (BioRad) to remove unincorporated label. Duplex substrates were prepared by heating equimolar concentrations of complementary strands to 95°C for 15 minutes followed by cooling to room temperature overnight and then purification on an S-300 spin column (GE) to remove single stranded DNA. Three duplex substrates – 5mC dsDNA (5mC DNA US and 5mC DNA LS), m^6A dsDNA (m^6A DNA US and m^6A DNA LS), and nm dsDNA (nm DNA US and nm DNA LS) – were prepared. Three single-stranded substrates – m^6A ssRNA 5mer, m^6A ssRNA 7mer, and m^6A ssDNA US – were left untreated and used accordingly. See Supplementary Table 1 for oligonucleotide sequences.

Filter binding assays

The standard buffer for the DNA and RNA binding assays contained 25 mM MES (pH 6.5), 2.0 mM $MgCl_2$, 0.1 mM DTT, 0.01 mM EDTA, and 40 μg/mL BSA. Binding was performed with purified SmMcrB 3-180 or mutants at 30°C for 10 min in a 30 μL reaction mixture containing 14.5 nM unlabeled DNA and 0.5 nM γ32P-labelled DNA. Samples were filtered through KOH-treated nitrocellulose filters (Whatman Protran BA 85, 0.45 μM) using a Hoefer FH225V filtration device for approximately 1 min. Filters were subsequently analyzed by scintillation counting on a 2910TR digital, liquid scintillation counter (PerkinElmer). All samples were measured in triplicate, averaged, and compared to a negative control to determine fraction bound.

Accession numbers

The atomic coordinates and structure factors for the N-terminal domain of *Staphylothermus marinus* McrB (residues 3-180) has been deposited in the Protein Data Bank (http://www.rcsb.org) under the PDB code 6N0S.

Acknowledgements

This work was supported National Institutes of Health Grant GM120242 (to J.S.C.) and is based upon research conducted at the Northeastern Collaborative Access Team (NE-CAT) beamlines under general user proposal GUP51113 (PI: J.S.C), which are funded by the National Institute of General Medical Sciences from the National Institutes of Health (P30 GM124165). The Pilatus 6M detector on 24-ID-C beam line is funded by a NIH-ORIP HEI grant (S10 RR029205). This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. J.S.C. is a Meinig Family Investigator in the Life Sciences. We thank the NE-CAT beamline staff for assistance with remote X-ray data collection.

Author Contributions

C.J.H. cloned, expressed, purified all Sm3-180 constructs and carried out filter binding experiments. C.J.H. and Y.N. crystallized Sm3-180, collected X-ray diffraction data, and solved the structure. C.J.H., Y.N., and J.S.C built the model and carried out computational modeling. C.J.H. and J.S.C. designed the study and wrote the manuscript.

Conflict of Interest

The authors declare that they have no competing financial interests.

REFERENCES

Adams, P.D., Afonine, P.V., Bunkóczi, G., Chen, V.B., Davis, I.W., Echols, N., *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr* **D66**:213-221.

Bertonati, C., Punta, M., Fischer, M., Yachdav, G., Forouhar, F., Zhou. W., *et al.* (2009) Structural genomics reveals EVE as a new ASCH/PUA-related domain. *Proteins* **75**:760-773.

Burman, R.W., Yates, P.A., Green, L.D., Jacky, P.B., Turker, M.S., and Popovich, B.W. (1999) Hypomethylation of an expanded FMR1 allele is not associated with a global DNA methylation defect. *Am J Hum Genet* **65**:1375-1386.

Chen, I.A., Markowitz, V.M., Chu, K., Palaniappan, K., Szeto, E., Pillay, M., *et al.* (2017) IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res* **45**:D507-D516.

Chotai, K.A., and Payne, S.J. (1998) A rapid, PCR based test for differential molecular diagnosis of Prader—Willi and Angelman syndromes. *J Med Genet* **35**:472-475.

Dryden, D., Murray, N.E., and Rao, D. (2001) Nucleoside triphosphate-dependent restriction enzymes. *Nucleic Acids Res* **29**:3728-3741.

Dupuis, M.È., Villion, M., Magadán, A.H., and Moineau, S. (2013) CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance. *Nat Commun* **4**:2087.

Ishikawa, K., Fukuda, E., and Kobayashi, I. (2010) Conflicts targeting epigenetic systems and their resolution by cell death: novel concepts for methyl-specific and other restriction systems. *DNA Res* **17**:325-342.

Iyer, L.M., Burroughs, A.M., and Aravind, L. (2006) The ASCH superfamily: novel domains with a fold related to the PUA domain and a potential role in RNA metabolism. *Bioinformatics* **22**:257–263.

Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L.E., *et al.* (2018) Improvements to the APBS biomolecular solvation software suite. *Protein Sci* **27**:112-128.

Kabsch, W. (2010) XDS. *Acta Crystallogr* **D66**:125-132.

Kim, B.N., Shin, M., Ha, S.C., Park, S.Y., Seo ,P.W., Hofmann, A., *et al.* (2017) Crystal structure of an ASCH protein from *Zymomonas mobilis* and its ribonuclease activity specific for single-stranded RNA. *Scientific Reports* **7**:12303.

Krüger, T., Wild, C., and Noyer-Weidner, M. (1995) McrB: a prokaryotic protein specifically recognizing DNA containing modified cytosine residues. *EMBO J* **14**: 2661-2669.

Labrie SJ, Samson JE, Moineau S (2010) Bacteriophage resistance mechanisms. Nat Rev Microbiol 8:317-327.

Li, F., Zhao, D., Wu, J., and Shi, Y. (2014) Structure of the YTH domain of human YTHDF2 in complex with an m(6)A mononucleotide reveals an aromatic cage for m(6)A recognition. *Cell Res* **24**:1490-1492.

Loenen, W.A., and Raleigh, E.A. (2014) The other face of restriction: modification-

dependent enzymes. *Nucleic Acids Res* **42**:56-69.

Luo, S., and Tong, L. (2014) Molecular basis for the recognition of methylated adenines in RNA by the eukaryotic YTH domain. *Proc Natl Acad Sci USA*. **111**:13834-13839.

O'Driscoll, J., Heiter, D.F., Wilson, G.G., Fitzgerald, G.F., Roberts, R., and Van Sinderen, D. (2006) A genetic dissection of the LlaJI restriction cassette reveals insights on a novel bacteriophage resistance system. *BMC Microbiol* **6**:40-52.

O'Sullivan, D.J., Zagula, K., and Klaenhammer, T.R. (1995) In vivo restriction by LlaI is encoded by three genes, arranged in an operon with llaIM, on the conjugative Lactococcus plasmid pTR2030. *J Bacteriol* **177**:134-143.

Ohshima, H., Matsuoka, S., Asai, K., and Sadaie, Y. (2002) Molecular organization of intrinsic restriction and modification genes BsuM of Bacillus subtilis Marburg. *J Bacteriol* **184**:381-399.

Panne, D., Raleigh, E.A., and Bickle, T.A. (1999) The McrBC endonuclease translocates DNA in a reaction dependent on GTP hydrolysis. *J Mol Biol* **290**:49-60.

Panne, D., Müller, S.A., Wirtz, S., Engel, A., and Bickle, T.A. (2001) The McrBC restriction endonuclease assembles into a ring structure in the presence of G nucleotides. *EMBO J* **20**:3210-3217.

Perez-Arellano, I., Gallego, J., and Cervera, J. (2007) The PUA domain - a structural and functional overview. *FEBS J* **274**:4972–4984.

Pieper, U., Schweitzer, T., Groll, D.H., Gast, F.U., and Pingoud, A. (1999) The GTP-binding domain of McrB: more than just a variation on a common theme? *J Mol Biol* **292**: 547-556.

Pieper, U., Pingoud, A. (2002) A mutational analysis of the PD...D/EXK motif suggests that McrC harbors the catalytic center for DNA cleavage by the GTP-dependent restriction enzyme McrBC from Escherichia coli. *Biochemistry* **41**:5236-5244.

Pieper, U., Groll, D.H., Wünsch, S., Gast, F.U., Speck, C., Mücke, N., et al. (2002) The GTP-dependent restriction enzyme McrBC from Escherichia coli forms high-molecular mass complexes with DNA and produces a cleavage pattern with a characteristic 10-base pair repeat. *Biochemistry* **41**:5245-5254.

Sheldrick, G.M. (2008) A short history of SHELX. *Acta Crystallogr* **A64**:112-122.

Stewart, F.J., Panne, D., Bickle, T.A., and Raleigh, E.A. (2000) Methyl-specific DNA binding by McrBC, a modification-dependent restriction enzyme. *J Mol Biol* **298**:611-622.

Sukackaite, R., Grazulis, S., Tamulaitis, G., and Siksnys, V. (2012) The recognition domain of the methyl-specific endonuclease McrBC flips out 5-methylcytosine. *Nucleic Acids Res* **40**:7552-7562.

Sutherland, E., Coe, L., and Raleigh, E.A. (1992) McrBC: a multisubunit GTP-dependent restriction endonuclease. *J Mol Biol* **225**:327-348.

Theler, D., Dominguez, C., Blatter, M., Boudet, J., Allain, F.H. (2014) Solution structure of the YTH domain in complex with N6-methyladenosine RNA: a reader of methylated RNA. *Nucleic Acids Res* **42**:13911-13919.

Tock, M.R., and Dryden, D.T. (2005) The biology of restriction and anti-restriction. *Curr Opin Microbiol* **8**:466-472.

Van Duyne, G.D., Standaert, R.F., Karplus, P.A., Schreiber, S.L., and Clardy, J. (1993) Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin. *J Mol Biol* **229**:105–124.

Weigele, P., and Raleigh, E.A. (2016) Biosynthesis and Function of Modified Bases in Bacteria and Their Viruses. *Chem Rev* **116**:12655-12687.

Wang, C., Zhu, Y., Bao, H., Jiang, Y., Xu, C., Wu, J., et al. (2016) A novel RNA-binding mode of the YTH domain reveals the mechanism for recognition of determinant of selective removal by Mmi1. *Nucleic Acids Res* **44**:969-982.

Xu, C., Wang, X., Liu, K., Roundtree, I.A., Tempel, W., Li Y, *et al.* (2014) Structural basis for selective binding of m6A RNA by the YTHDC1 YTH domain. *Nat Chem Biol* **10**:927-929.

Xu, C., Liu. K., Ahmed, H., Loppnau, P., Schapira, M., and Min, J. (2015) Structural Basis for the discriminative recognition of N6-methyladenosine RNA by the human YT521-B homology domain family of proteins. *J Biol Chem* **290**:24902-24913.

Zhu, T., Roundtree, I.A., Wang, P., Wang, X., Wang, L., Sun, C., et al. (2014) Crystal structure of the YTH domain of YTHDF2 reveals mechanism for recognition of N6-methyladenosine. *Cell Res* **24**:1493-1496.

Table 1. X-ray data collection and refinement statistics for SmMcrB 3-180

Data collection	
PDB code	6N0S
X-ray Source	NECAT 24-ID-C
Wavelength (Å)	0.9791
Space group	$P4_32_12$
Unit cell	a = 62.41, b = 62.41, c = 118.63 Å α = 90.00°, β = 90.00°, γ = 90.00°
Resolution, Å[a]	55.26 – 2.10 (2.34 – 2.10)
No. measured reflections[a]	1297122 (171896)
No. unique reflections[a]	61793 (7394)
Completeness (%)[a]	99.9 (100.0)
Multiplicity[a]	42.2 (23.2)
R_{merge}[a]	0.113 (0.429)
Mean I/σ_I[a]	12.6 (8.22)
$CC_{1/2}$[a]	0.999 (0.993)
Refinement	
R_{work}/R_{free}	0.1880 / 0.2298
RMSD	
Bond lengths (Å)	0.013
Bond angles (°)	1.021
Ramachandran plot	
Favored (%)	96.09
Allowed (%)	3.91
Outliers (%)	0.00
Average B-Factor	45.04
Clashscore	4.23
No. Atoms	
Macromolecule	1522
Solvent	94
Sulfate	10

[a] Parentheses indicate values for highest resolution shell

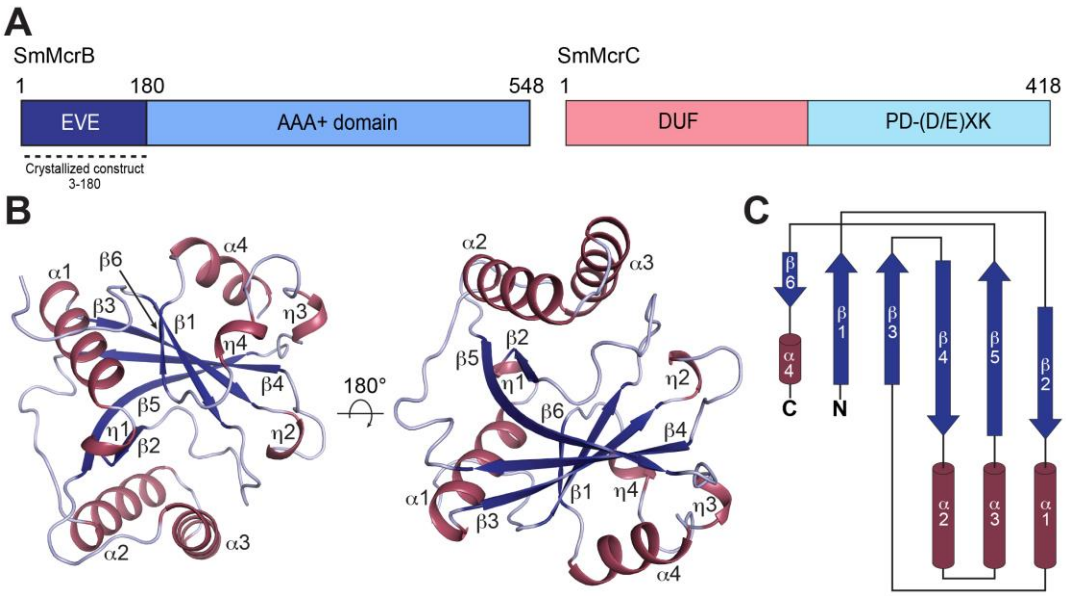

Figure 1. Structure and topology of Sm3-180.

A. Domain architecture of SmMcrBC. Dashed line denotes crystallized construct (residues 3-180). B. Cartoon representations of SmMcrB 3-180 in two orientations. Helices and β-strands are colored raspberry and blue respectively. C. Topology diagram of SmMcrB 3-180.

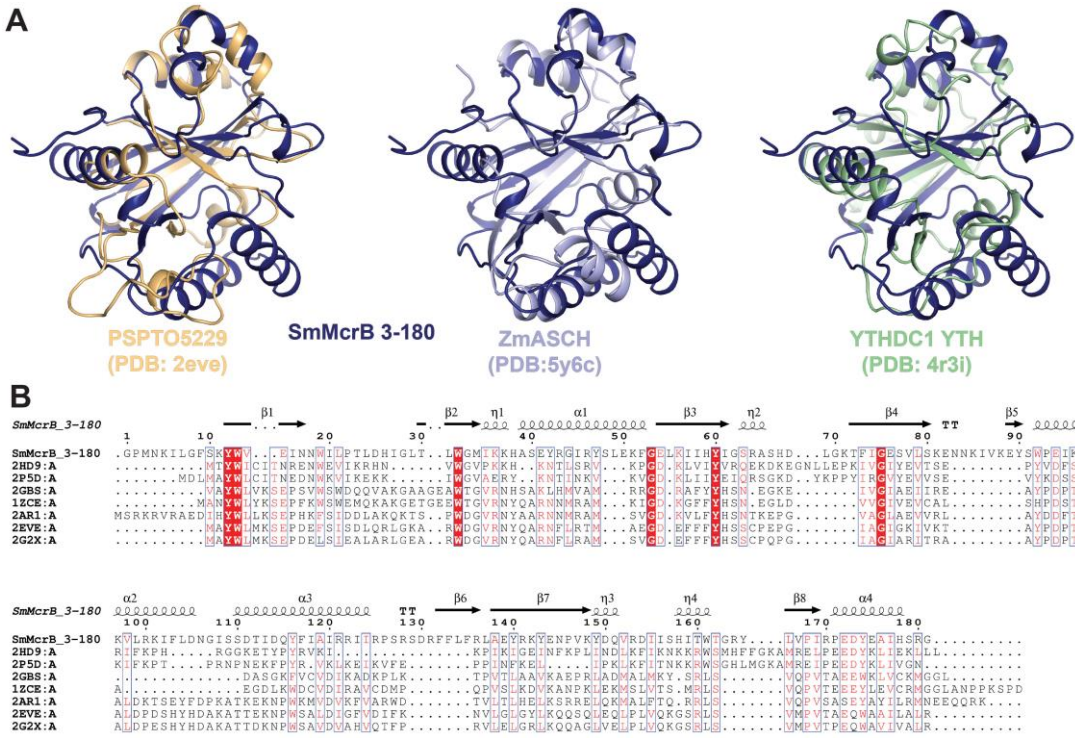

Figure 2. Sm3-180 adopts an EVE domain fold.

A. Structural superpositions of SmMcrB 3-180 with PSPTO5229 (PDB: 2eve, Z-score = 9.8, RMSD = 2.9Å), ZmASCH (PDB: 5y6c, Z-score = 10.9, RMSD = 2.6Å), and YTHDC1 YTH (PDB: 4r3i, Z-score = 9.1, RMSD = 2.6Å). B. Structure based sequence alignment of SmMcrB 3-180 with EVE domain homologs. Secondary structure of SmMcrB 3-180 is mapped above alignment. Conserved residues that distinguish EVE domains from YTH domains are present in SmMcrB 3-180 and colored red. Sequence labeling associated with the listed PDB codes is as follows: 2HD9:A, PH1033 from *Pyrococcus horikoshii OT3*; 2P5D:A, MJECL36 from *Methanocaldococcus jannaschii DSM 2661*; 2GBS:A, Rpa0253 from *Rhodopseudomonas palustris*; 1ZCE:A, Atu2648 from *Agrobacterium tumefaciens*; 2AR1:A, Hypothetical protein from *Leishmania major*; 2EVE:A, PSPTO5229 from *Pseudomonas syringae*; 2G2X:A, Q88CH6 from *Pseudomonas putida*.

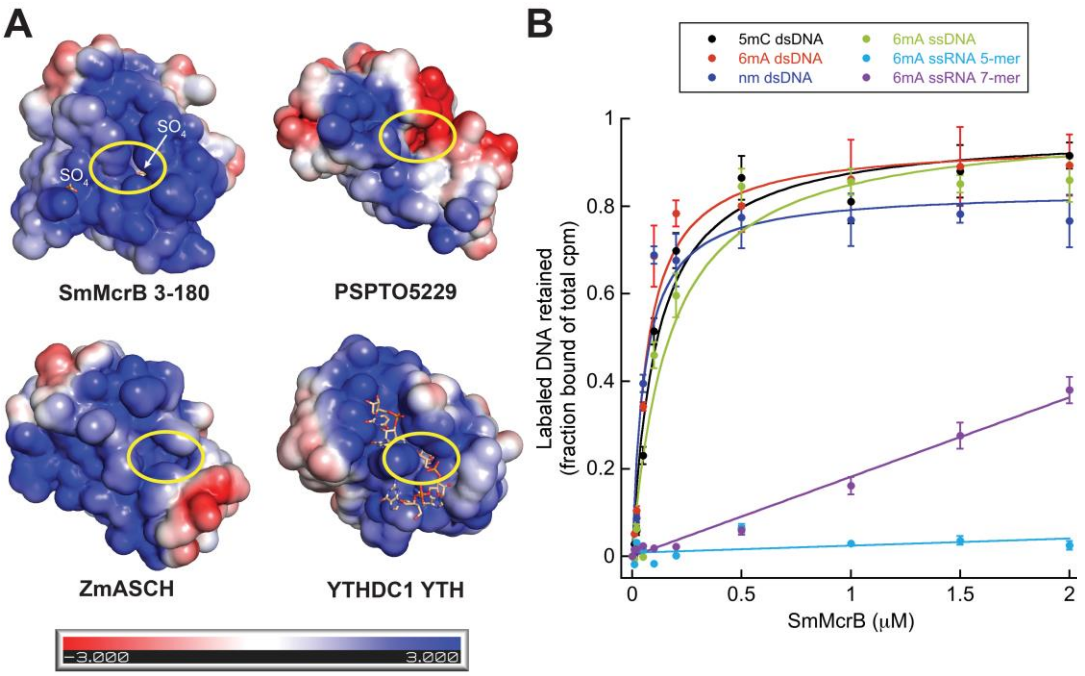

Figure 3. SmMcrB preferentially binds DNA and is promiscuous.

A. Electrostatic surfaces of SmMcrB 3-180, PSPTO5229 (PDB: 2eve), ZmASCH (PDB: 5y6c), and YTHDC1 YTH domain with bound 5'-GG(m^6A)CU-3' ssRNA (PDB: 4r3i). Bound sulfate ions and RNA are shown in stick representation and colored yellow and wheat respectively. Yellow circle shows location of the aromatic cage binding pocket. Scale bar indicates electrostatic surface coloring from -3 K_bT/e_c to +3 K_bT/e_c. B. Filter binding of SmMcrB 3-180 with DNA and RNA shows preference for DNA.

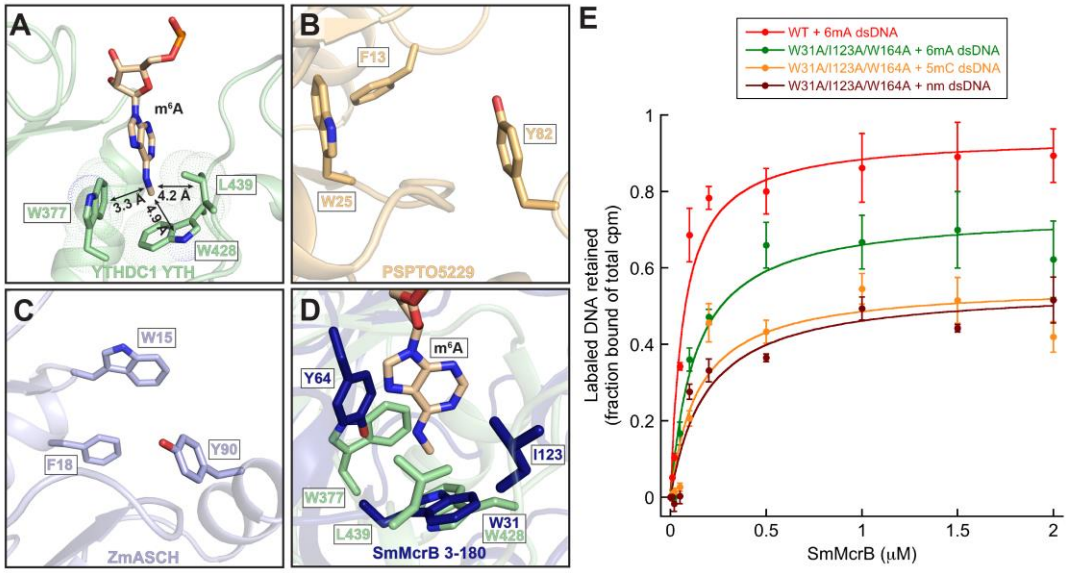

Figure 4. SmMcrB contains a degenerate aromatic cage.

A. Zoomed view of aromatic cage from YTHDC1 with bound m^{6}A (PDB: 4r3i). Cage residues are highlighted with surface dots and labeled. m^{6}A base from bound RNA is shown (wheat). Calculated interaction distances are indicated. B. Zoomed view of PSPTO5229 aromatic cage (light orange; PDB: 2eve). C. Zoomed view of ZmASCH aromatic cage (light blue, PDB: 5y6c). D. Superposition of RNA-bound YTHDC1 YTH (light green) with SmMcrB 3-180 (dark blue). Cage residues from each structure are shown as sticks and labeled. m^{6}A base from RNA substrate is colored wheat. E. Filter binding of SmMcrB 3-180 WT with m^{6}A dsDNA (red) and W31A/I123A/W164A triple mutant with m^{6}A dsDNA (green), 5mC dsDNA (orange), and non-methylated (nm) dsDNA (brown).

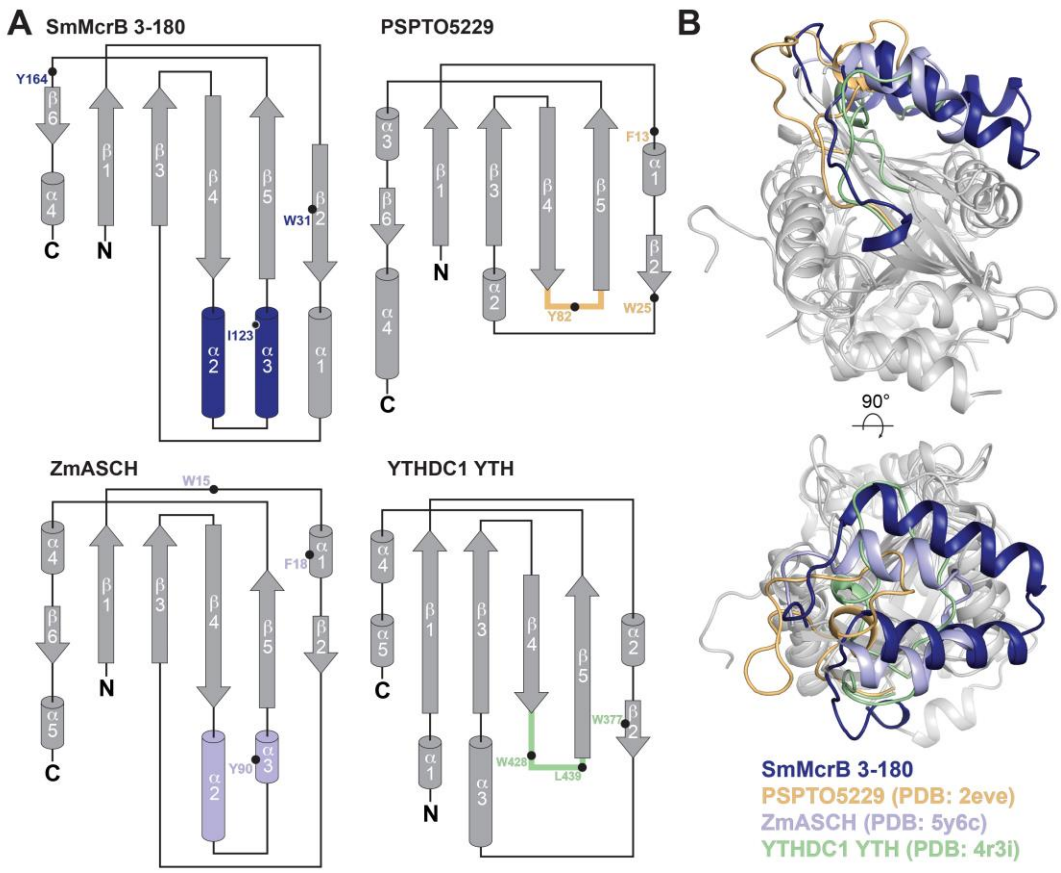

Figure 5. Organization of structural motifs dictating binding specificity in PUA-like domains.

A. Topology diagrams of the of PUA-like domains. Structural core is colored gray with respective β4-β5 loops and inserts are colored as follows: SmMcrB 3-180, dark blue; PSPTO5229, light orange; ZmASCH, light blue; YTHDC1, light green. Relative positions of aromatic cage residues are marked with black circles and labeled. B. Structural superposition of PUA like domains highlighting differences in the β4-β5 loop. Structural motifs colored as in A.

Chapter 5. Concluding Remarks and Future Directions

Concluding Remarks

My work described here aims to redefine the modification dependent restriction system, McrBC. McrBC has historically been identified as a modified cytosine restriction system. Bioinformatic analysis coupled with structural studies performed here show that not only are the N-terminal DNA binding modules highly divergent amongst McrB homologs, but that they have evolved to have varying target specificities and modalities of recognition. For instance, the *Helicobacter pylori* LlajI.R1 N-terminal domain adopts a B3 domain for site-specific recognition, the *Thermococcus gammatolerans* McrB N-terminal domain adopts a YTH domain for m^6A DNA recognition, and the *Staphylothermus marinus* McrB N-terminal domain adopts an EVE domain for nonspecific DNA recognition.

Interestingly, the remainder of the McrBC system is highly conserved, including the McrB C-terminal GTP-specific AAA+ and the McrC PD-(D/E)XK endonuclease. In fact, in *E. coli*, the C-terminal domain alone is capable of reconstituting stimulated GTP hydrolysis and the only thing preventing it from being functional on DNA is the lack of a DNA binding module. McrBC therefore seems to be a highly modular restriction system that has incorporated different N-terminal domains to target the system to varying sites. The selective pressure behind this evolution is presumably to protect itself from phages with different signatures to escape restriction. This unique modularity provides a potential platform for the engineering of novel restriction enzymes with potentially limitless recognition sites.

Future Directions

Although my work has aimed to redefine the McrBC system as a modular restriction enzyme with a broad scope of targets, many other McrB homologs with varying N-terminal domains still exist. Furthermore, many questions regarding its mechanism of action in GTP binding, stimulated hydrolysis, and cleavage remain. A crystal structure of the AAA+ domain from an McrB homolog, LlaI.2, solved in our lab is the first stepping stone to answering some of these more complex questions. The structure is briefly described in Appendix 1 and revealed some surprising results. These observations are corroborated by the crystal structure of the TgMcrB AAA+ domain also solved by our lab (not discussed here). Together these structures have identified key catalytic residues involved in GTP binding and GTP hydrolysis. The table below lists some of these residues that are potential targets for mutagenesis in our functional assays. Structural conservation between LlaI.2 and TgMcrB AAA+ allowed us to also identify the analogous residues in EcMcrB.

Although simple GTP hydrolysis/stimulation experiments of these mutants would prove useful, two additional key experiments would provide the necessary information to put structure and function in the context of bacterial defense and phage restriction. These are 1) *in vivo* phage competition assays with wild-type and mutant McrBC systems and 2) *in vitro* cleavage assay with substrates varying in m^5C spacing. The *in vivo* results (coupled with *in vitro* GTP hydrolysis) would irrefutably allow us to argue the residues observed in our structure are responsible for stimulated GTP hydrolysis. These would be measured by cell viability in the presence and absence of phage harboring m^5C DNA. The *in vitro* cleavage assay would allow us to measure the

TgMcrB	EcMcrB	Function	position
K221	K207	Walker A p-loop	cis
T222	T208	Mg2+ binding	cis
W223	F209	guanine binding	cis
D356	D279	Walker B, mg2+ binding	cis
E357	E280	Walker B, mg2+ binding	cis
E375	E298	ribose interacting	trans
D377	D300	ribose, ab-phosphate interacting	trans
N410	N333	g4 MNTAD, positions catalytic water	cis
D413	D336	g4 MNTAD, positions N410 (N333)	cis
D420	D343	positions D413 (D336)	trans
R425	R348	Arg from sensor 2	trans
R426	R349	Arg finger	trans
H501	H407	guanine binding	cis
TgMcrC	EcMcrC	Function	position
R263	K157	positions D413 (D336)	
	D244	Mg2+ binding	
	D257	Mg2+ binding	
	K259	charge compensation	

dependence of stimulated GTP hydrolysis on translocation of McrBC. In an ideally spaced substrate of 40-60 bp, McrBC should be functional in GTPase dead mutants as translocation is not necessary. Likewise, in a long-ranged substrate (100-2000 bp), McrBC should not be functional in a GTPase dead mutant as translocation is necessary.

The final key experiment necessary to bring this project full circle is to solve the crystal or cryo-EM structure of the full-length McrBC complex + DNA. It cannot be understated the value of an atomic resolution structure of this assembly, especially if captured in a cleavage competent conformation. This structure would aim to reveal 1) the organization of the DNA binding domains around the McrB oligomer, 2) the interaction of McrB with McrC, and 3) the organization of DNA on the complex, particularly at the DNA binding sites and the endonuclease active site.

Appendix 1. Crystal Structure of LlaI.2

McrBC is a two-component modification dependent restriction that includes the McrB and McrC proteins. McrB is a two-domain protein with an N-terminal putative DNA binding domain and a C-terminal GTP-specific AAA+ domain. McrC is also a two domain protein with an N-terminal DUF and a C-terminal PD-(D/E)XK endonuclease domain. Previous studies show that McrB undergoes GTP-dependent oligomerization and has a low basal level of GTP hydrolysis that is stimulated by the presence of McrC (Pieper et al., 1997 and 1999). To understand how this unique AAA+ domain is capable of binding and hydrolyzing GTP, we have solved the crystal structure of an McrB homolog, LlaI.2, in the presence of GDP-AlF_x to 1.92 Å. LlaI.2 is part of the LlaI restriction/modification cassette in *Lactococcus lactis*. This cassette encodes for four proteins, M1 – a methyltransferase, R1 – a putative DNA binding domain, R2 – a GTP specific AAA+, and R3 – a PD-(D/E)XK endonuclease. The fact that LlaI.2 exists as its own independent AAA+ was a deciding factor in choosing it as a potential crystallization target.

KEY RESULTS AND DISCUSSION

Crystals of LlaI.2 were grown in the presence of the GTP transition state mimic, GDP-AlF_x, and data was collected at the NE-CAT 24-ID-C beamline (Figure 1). The crystal structure of LlaI.2 reveals that it oligomerizes as a hexamer in the space group H3 (Figure 2). Although there are only two molecules in the asymmetric unit, multiple symmetry related molecules were used to create the full hexamer. A recent study on the *E. coli* McrB revealed that it also oligomerizes as a hexamer in the presence of GTP and suggests this organization may be conserved across McrB homologs (Nirwan, et al.,

2019).

GDP was bound in every other molecule within the hexameric arrangement with a citrate from the crystallization condition bound in every adjacent molecule (Figure 2). Inspection of the GDP bound molecules revealed that guanine base recognition is achieved via π-stacking with F235 from above the indole plane and R17 from below. An additional glutamine Q174 hydrogens bonds with the carbonyl at the 6 position of the guanine base. The R17 residue immediately succeeds the P-loop and an aromatic residue is conserved in this position for all McrB homologs (Figure 4). This mode of GTP binding is unique among all GTPases that canonically utilize a conserved aspartate or glutamate to achieve guanine specificity at the 1 and 2 positions, as seen in Ras and $G_{i\alpha1}$ (Scheffzek, et al., 1997; Tesmer, et al., 1997).

Interestingly, the conserved McrB signature sequence (NTAD, discussed in Chapter 1) does not associate with the guanine base. It is instead poised similarly to the conserved glutamine from switch II used to position the catalytic water, as seen in Ras-RasGAP (Figure 5). Although the protein was crystallized in the presence of GDP-AlF_x, AlF_x, and thus the catalytic water, were not bound and could not be modeled. This is a direct consequence of the adjacent molecules bound to citrate exerting a conformational change to the GDP binding site, precluding GDP, AlF_x, and water from binding (Figure 6). This also resulted in a major, downward shift of the conserved Mg^{2+} binding residues S16 from Walker A and E88 and E89 from Walker B from the phosphates, also precluding Mg^{2+}. Finally, catalytic residues normally positioned *in trans*, including the arginine finger, were over 10 Å away from the active site and could not be identified.

Although the crystal structure of LlaI.2 failed to provide a comprehensive view

of the GTP active site, it did provide evidence for several key observations. First, LlaI.2 (and EcMcrB) are definitively hexameric. Second, LlaI.2 utilizes a unique GTP binding mode via π-stacking and reading of the 6 position. Third, the conserved NTAD signature sequence appears to be involved in positioning of the catalytic water, not for guanine specificity as previously described (Pieper et al., 1997 and 1999). Despite these findings, further investigation is warranted to fully elucidate the mechanisms of GTP binding and hydrolysis by McrB homologs.

EXPERIMENTAL PROCEDURES

Cloning, expression and purification of LlaI.2

DNA encoding the *Lactococcus lactis* LlaI.2 protein was codon optimized for *E. coli* expression and synthesized commercially from Integrated DNA Technologies (IDT), Inc. DNA encoding the full-length LlaI.2 (residues 1-336) was amplified by PCR and cloned into pET21b, introducing a 6xHis tag at the C-terminus. Native LlaI.2 was transformed into BL21(DE3) cells, grown at 37°C in Terrific Broth to an OD600 of 1.0, and then induced with 0.3 mM IPTG overnight at 19°C. SeMet LlaI.2 was expressed in minimal media using methionine auxotrophs (T7 Express Crystal Competent E. coli, New England Biolabs) according to manufacturer protocols. The cells were harvested and washed twice with a nickel load buffer (20 mM HEPES pH 7.5, 500 mM NaCl, 30 mM Imidazole, 5% glycerol (v/v), and 5 mM bME). Pellets were typically flash frozen in liquid nitrogen and stored at -80°C. Thawed pellets from 500 mL cultures were resuspended in 30 mL of a nickel load buffer supplemented with 10 mM PMSF, 5 mg DNase I (Roche), 5 mM $MgCl_2$ and a complete protease inhibitor cocktail tablet

(Roche). Lysozyme was added to 1 mg/mL and the mixture was incubated for 15 minutes with rocking at 4°C. Cells were disrupted by sonication and the lysate was cleared of debris by centrifugation at 13,000 rpm (19,685x g) for 30 minutes at 4°C. The supernatant was filtered, loaded onto a 5-ml HiTrap chelating column charged with $NiSO_4$ and then washed with a nickel load buffer. LlaI.2 was eluted with an imidazole gradient from 30 mM to 1 M. Pooled fractions were concentrated on a 30 kDa Millipore centrifugal concentrator. The concentrated protein was further purified by size exclusion chromatography (SEC) using a Superdex 200 16/600 pg column. All proteins were exchanged into a final buffer of 20 mM HEPES pH 7.5, 2M NaCl, 5 mM MgCl2, and 1 mM DTT (5 mM for SeMet labelled) during SEC and concentrated to 30-60 mg/ml. Concentrations of purified proteins were determined by SDS-PAGE with BSA standards.

Crystallization, X-ray data collection and structure determination of LlaI.2

LlaI.2 (12 – 20 mg/mL) were prepared with GDP-AlF_X (2.5 mM GDP, 25 mM NaF and 2.5 mM $AlCl_3$). Both native and SeMet LlaI.2 were crystallized by sitting drop vapor diffusion in 0.2 M tri-ammonium citrate pH 6.0, 20% PEG 6000 with a drop size of 2 μL and reservoir volume of 65 μL. Crystals typically appeared within 1-2 weeks at 20°C and were cryoprotected with Parabar 10312 from Hampton Research and frozen in liquid nitrogen. They were of the space group H3 with unit cell dimensions a = 177.75 Å, b = 177.75 Å, c = 65.03 Å and $\alpha = 90.00°$, $\beta = 90.00°$, $\gamma = 120.00°$. Single-wavelength anomalous diffraction (SAD) data were collected remotely on the tunable NE-CAT 24-ID-C beamline at the Advanced Photon Source at the selenium edge energy at 12.663

keV (0.9791 Å) (Table 1). Data were integrated and scaled using XDS (Kabsch, 2010) and AIMLESS (Evans, 2006) via the NE-CAT RAPD pipeline. Heavy atom sites were located using SHELX (Sheldrick, 2008) and phasing, density modification, and initial model building was carried out using the Autobuild routines of the PHENIX package (Adams et al., 2010). Further model building and refinement was carried out manually in COOT (Emsley et al., 2010) and PHENIX (Adams et al., 2010) respectively. The final model was refined to 1.92Å resolution with Rwork/Rfree = 0.1830/0.2109 (Table 1) and contained two molecules in the asymmetric unit.

REFERENCES

Nirwan N, Singh P, Mishra GG, Johnson CM, Szczelkun MD, Inoue K, Vinothkumar KR, Saikrishnan K. Hexameric assembly of the AAA+ protein McrB is necessary for GTPase activity. Nucleic Acids Res. 2019 Jan 25;47(2):868-882.

Pieper U, Brinkmann T, Krüger T, Noyer-Weidner M, Pingoud A. Characterization of the interaction between the restriction endonuclease McrBC from E. coli and its cofactor GTP. J Mol Biol. 1997 Sep 19;272(2):190-9.

Pieper U, Schweitzer T, Groll DH, Gast FU, Pingoud A. The GTP-binding domain of McrB: more than just a variation on a common theme? J Mol Biol. 1999 Sep 24;292(3):547-56.

Scheffzek K, Ahmadian MR, Kabsch W, Wiesmüller L, Lautwein A, Schmitz F, Wittinghofer A. The Ras-RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants. Science. 1997 Jul 18;277(5324):333-8.

Tesmer JJ, Berman DM, Gilman AG, Sprang SR. Structure of RGS4 bound to AlF4--

activated G(i alpha1): stabilization of the transition state for GTP hydrolysis. Cell. 1997 Apr 18;89(2):251-61.

Table 1. X-ray data collection and refinement statistics for LlaI.2

Data collection	
PDB code	
X-ray Source	NECAT 24-ID-C
Wavelength (Å)	0.9791
Space group	H3
Unit cell	a = 177.75, b = 177.75, c = 65.03 Å α = 90.00°, β = 90.00°, γ = 120.00°
Resolution, Å[a]	88.87 – 1.92 (2.07 – 1.92)
No. measured reflections[a]	1073325 (153257)
No. unique reflections[a]	84745 (24061)
Completeness (%)[a]	99.1 (100.0)
Multiplicity[a]	12.7 (6.36)
R_{merge}[a]	0.058 (0.521)
Mean I/σ_I[a]	21.8 (3.79)
$CC_{1/2}$[a]	1.000 (90.8)
Refinement	
R_{work}/R_{free}	0.1830 / 0.2109
RMSD	
Bond lengths (Å)	0.007
Bond angles (°)	0.911
Ramachandran plot	
Favored (%)	96.32
Allowed (%)	2.76
Outliers (%)	0.92
Average B-Factor	41.23
Clashscore	4.73
No. Atoms	
Macromolecule	5284
Solvent	367
Total	5651

[a] Parentheses indicate values for highest resolution shell

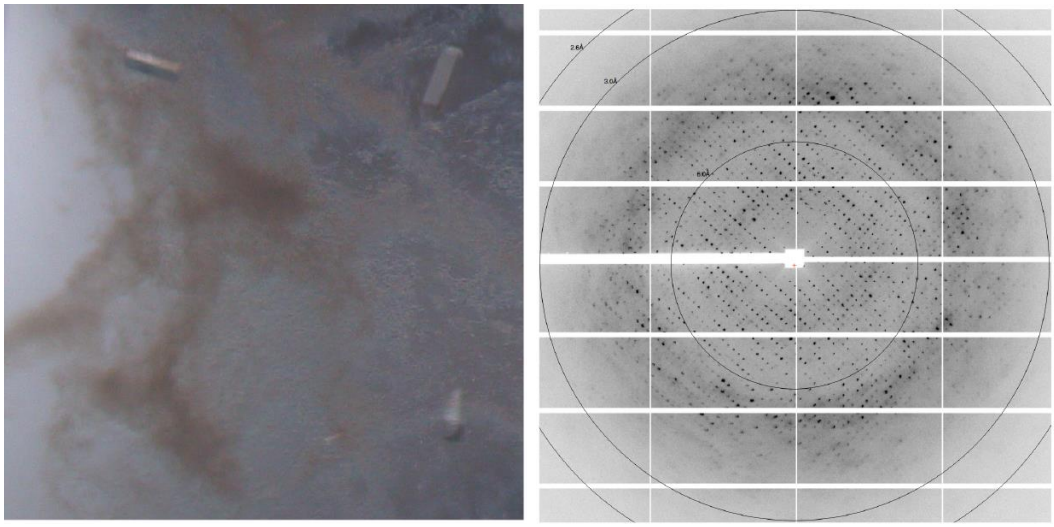

Figure 1. Crystallization and Data Collection of LlaI.2 + GDP-AlF$_x$. *Left.* Crystals of LlaI.2 + GDP-AlF$_x$. *Right.* Diffraction data of LlaI.2 + GDP-AlF$_x$ crystals collected at the NE-CAT 24-ID-C beamline at the Advanced Photon Source.

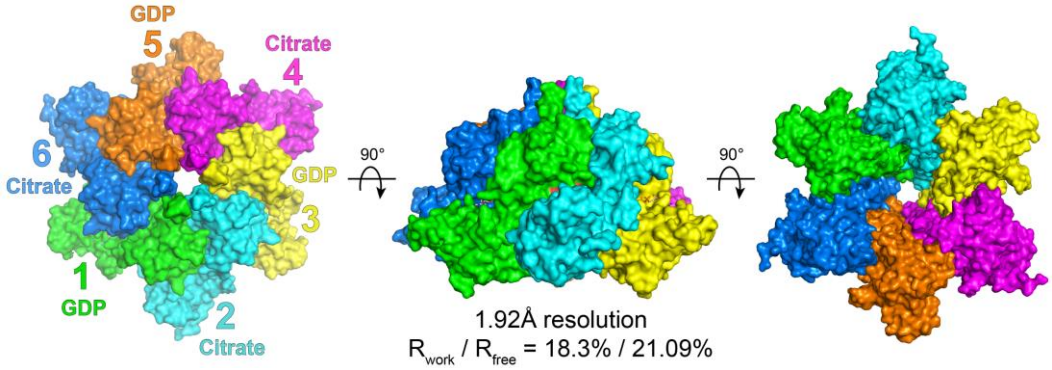

Figure 2. Surface representation of LlaI.2 + GDP-AlF$_x$. The structure of LlaI.2 + GDP-AlFx adopts a hexamer and is shown in three orientations. Each subunit within the hexamer is colored differently. Every odd molecule is bound to GDP while every even molecule is bound to citrate.

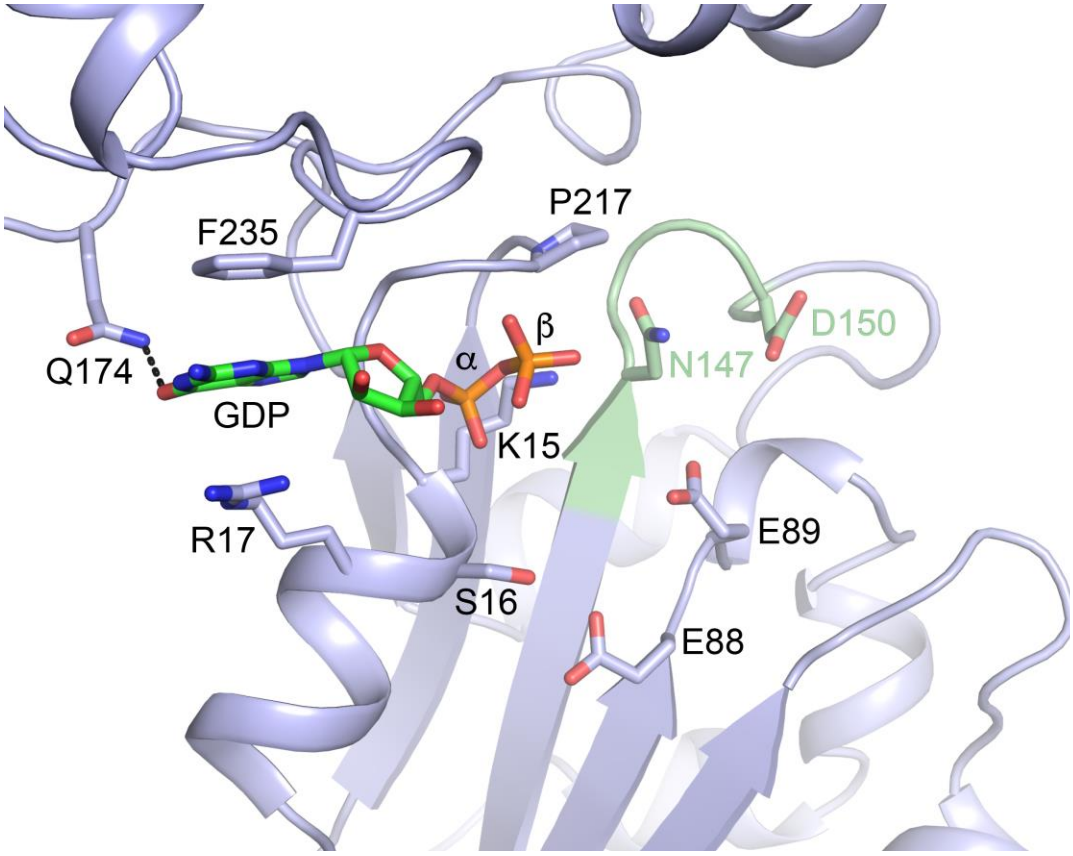

Figure 3. GDP binding site and conserved NxxD motif in LlaI.2. Zoomed in view of the GDP binding site in LlaI.2. The bound guanine base is stabilized via π-stacking with F235 and R17 while Q174 reads the carbonyl at the 6 position. The Walker A proline (P217), lysine (K15), and serine (S16) and the Walker B acidic residues (E88 and E89) are also shown in sticks and colored grey. The conserved McrB NxxD motif is shown in sticks and colored green (N147 and D150).

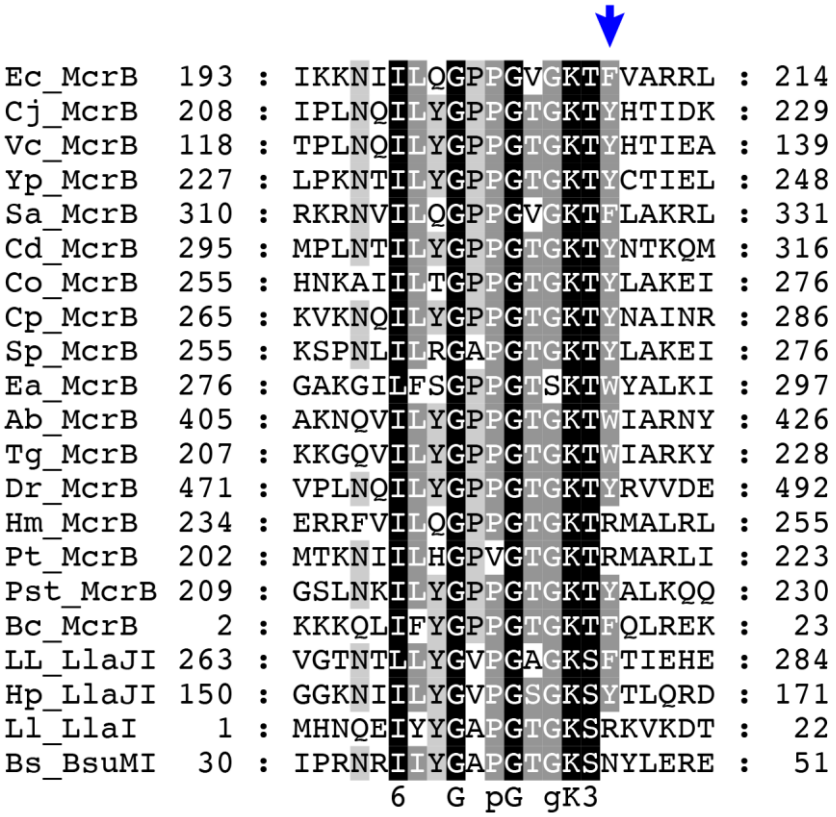

Figure 4. Sequence alignment of McrB Walker A motifs. The blue arrow denotes a conserved aromatic residue in the residue immediately succeeding the P-loop that is responsible for π-stacking with the guanine base.

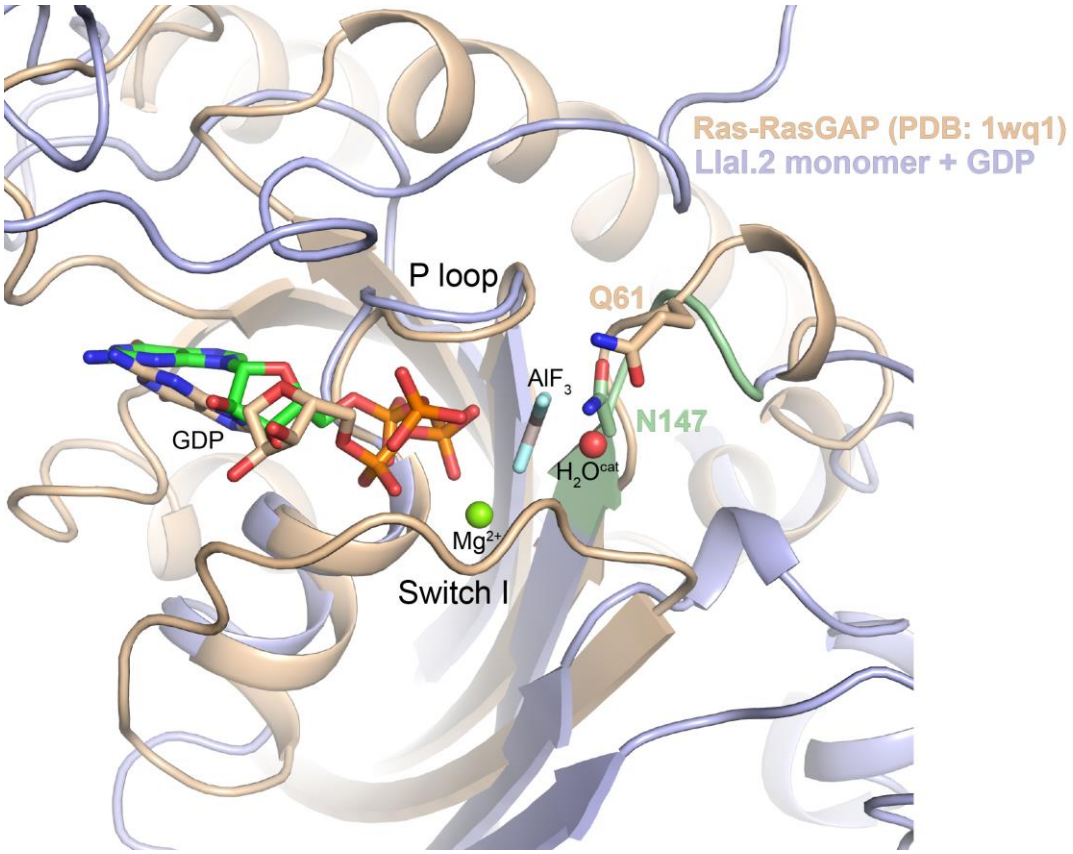

Figure 5. Structural superposition of LlaI.2 + GDP-AlF$_x$ with Ras-RasGAP. Structural superposition of LlaI.2 with Ras-RasGAP reveals structural similarity between the two GTPase active sites. Notably the Q61 from the switch II motif of Ras-RasGAP is positioned similarly as N147 from the NxxD of LlaI.2 and suggests it may be involved in positioning the catalytic water.

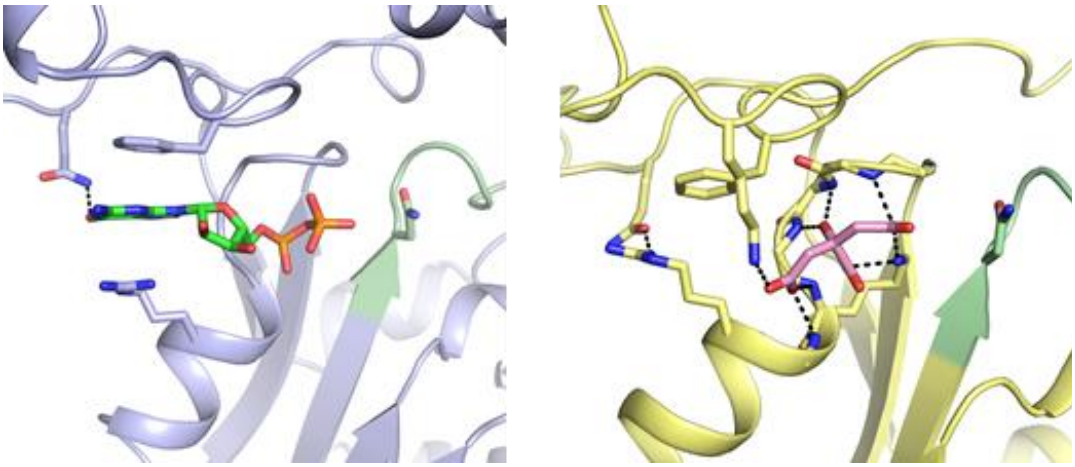

Figure 6. Structural comparison of LlaI.2 molecules bound with GDP or Citrate. *Left.* Zoomed in view of the LlaI.2 GTPase active site bound to GDP (green). *Right.* Zoomed in view of the LlaI.2 GTPase active site bound to citrate (pink). Citrate acts as a pyrophosphate mimetic and closes the active site and precludes GDP from every even molecule within the hexamer.